The ART of ANALOG LAYOUT

Alan Hastings

PRENTICE HALL
Upper Saddle River, NJ 07458

VISIT…

Library of Congress Cataloging-in-Publication Data
Hastings, Alan (Ray Alan)
The art of analog layout / Alan Hastings.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-087061-7
1. Integrated circuits—Design and construciton. 2. Layout (Printing) I. Title.

TK7874.H3926 2001
621.3815—dc21 00-045307

Vice president and editorial director, ECS: **Marcia Horton**
Publisher: **Tom Robbins**
Associate editor: **Alice Dworkin**
Editorial assistant: **Jessica Power**
Production editor: **Carlisle Communications, Ltd.**
Executive managing editor: **Vince O'Brien**
Managing editor: **David A. George**
Art director: **Jayne Conte**
Cover design: **Joseph Sengotta**
Art editor: **Adam Velthaus**
Manufacturing manager: **Trudy Pisciotti**
Manufacturing buyer: **Dawn Murrin**
Assistant vice president of production and manufacturing, ESM: **David W. Riccardi**

Prentice Hall

Copyright © 2001 by Prentice-Hall, Inc.
Upper Saddle River, New Jersey 07458.

All rights reserved. No part of this book may be
reproduced, in any form or by any means,
without the permission in writing from the publisher.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to the documentation contained in this book.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-087061-7

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Pearson Education Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

For My Father

Contents

Preface xvii
Acknowledgments xix

1 *Device Physics*
1.1 Semiconductors 1
1.1.1 Generation and Recombination 4
1.1.2 Extrinsic Semiconductors 6
1.1.3 Diffusion and Drift 9
1.2 PN Junctions 10
1.2.1 Depletion Regions 10
1.2.2 PN Diodes 13
1.2.3 Schottky Diodes 15
1.2.4 Zener Diodes 17
1.2.5 Ohmic Contacts 19
1.3 Bipolar Junction Transistors 20
1.3.1 Beta 22
1.3.2 I-V Characteristics 23
1.4 MOS Transistors 24
1.4.1 Threshold Voltage 27
1.4.2 I-V Characteristics 29
1.5 JFET Transistors 31
1.6 Summary 33
1.7 Exercises 34

2 *Semiconductor Fabrication*
2.1 Silicon Manufacture 36
2.1.1 Crystal Growth 37
2.1.2 Wafer Manufacturing 38
2.1.3 The Crystal Structure of Silicon 38
2.2 Photolithography 40
2.2.1 Photoresists 40
2.2.2 Photomasks and Reticles 41
2.2.3 Patterning 42
2.3 Oxide Growth and Removal 42
2.3.1 Oxide Growth and Deposition 43
2.3.2 Oxide Removal 44
2.3.3 Other Effects of Oxide Growth and Removal 46
2.3.4 Local Oxidation of Silicon (LOCOS) 48

2.4 Diffusion and Ion Implantation 49
2.4.1 Diffusion 50
2.4.2 Other Effects of Diffusion 52
2.4.3 Ion Implantation 53
2.5 Silicon Deposition 55
2.5.1 Epitaxy 56
2.5.2 Polysilicon Deposition 58
2.6 Metallization 58
2.6.1 Deposition and Removal of Aluminum 59
2.6.2 Refractory Barrier Metal 60
2.6.3 Silicidation 62
2.6.4 Interlevel Oxide, Interlevel Nitride, and Protective Overcoat 63
2.7 Assembly 64
2.7.1 Mount and Bond 66
2.7.2 Packaging 69
2.8 Summary 69
2.9 Exercises 69

3 *Representative Processes*
3.1 Standard Bipolar 72
3.1.1 Essential Features 72
3.1.2 Fabrication Sequence 73
Starting Material 73
N-Buried Layer 73
Epitaxial Growth 74
Isolation Diffusion 74
Deep-N+ 74
Base Implant 75
Emitter Diffusion 75
Contact 76
Metallization 76
Protective Overcoat 77
3.1.3 Available Devices 77
NPN Transistors 77
PNP Transistors 79
Resistors 81
Capacitors 83
3.1.4 Process Extensions 84
Up-down Isolation 84
Double-level Metal 84
Schottky Diodes 85
High-Sheet Resistors 86
Super-beta Transistors 86
3.2 Polysilicon-Gate CMOS 87
3.2.1 Essential Features 88

3.2.2 Fabrication Sequence 89
Starting Material 89
Epitaxial Growth 89
N-well Diffusion 89
Inverse Moat 90
Channel Stop Implants 90
LOCOS Processing and Dummy Gate Oxidation 91
Threshold Adjust 92
Polysilicon Deposition and Patterning 93
Source/Drain Implants 93
Contacts 94
Metallization 94
Protective Overcoat 94
3.2.3 Available Devices 95
NMOS Transistors 95
PMOS Transistors 97
Substrate PNP Transistors 98
Resistors 98
Capacitors 100
3.2.4 Process Extensions 100
Double-level Metal 100
Silicidation 101
Lightly Doped Drain (LDD) Transistors 101
Extended-Drain, High-Voltage Transistors 103
3.3 Analog BiCMOS 104
3.3.1 Essential Features 104
3.3.2 Fabrication Sequence 106
Starting Material 106
N-buried Layer 106
Epitaxial Growth 106
N-well Diffusion and Deep-N+ 107
Base Implant 107
Inverse Moat 108
Channel Stop Implants 108
LOCOS Processing and Dummy Gate Oxidation 108
Threshold Adjust 109
Polysilicon Deposition and Pattern 109
Source/Drain Implants 109
Metallization and Protective Overcoat 110
Process Comparison 110
3.3.3 Available Devices 111
NPN Transistors 112
PNP Transistors 112
Resistors 115
3.4 Summary 115
3.5 Exercises 116

4 *Failure Mechanisms*
4.1 Electrical Overstress 118
4.1.1 Electrostatic Discharge (ESD) 118
Effects 120
Preventative Measures 120
4.1.2 Electromigration 121
Effects 121
Preventative Measures 122
4.1.3 The Antenna Effect 122
4.2 Contamination 124
4.2.1 Dry Corrosion 124
Effects 124
Preventative Measures 125
4.2.2 Mobile Ion Contamination 125
Effects 125
Preventative Measures 126
4.3 Surface Effects 128
4.3.1 Hot Carrier Injection 128
Effects 128
Preventative Measures 130
4.3.2 Parasitic Channels and Charge Spreading 131
Effects 131
Preventative Measures (Standard Bipolar) 133
Preventative Measures (CMOS and BiCMOS) 137
4.4 Parasitics 139
4.4.1 Substrate Debiasing 140
Effects 140
Preventative Measures 142
4.4.2 Minority-Carrier Injection 143
Effects 143
Preventative Measures (Substrate Injection) 146
Preventative Measures (Cross-injection) 151
4.5 Summary 153
4.6 Exercises 153

5 *Resistors*
5.1 Resistivity and Sheet Resistance 156
5.2 Resistor Layout 158
5.3 Resistor Variability 162
5.3.1 Process Variation 162
5.3.2 Temperature Variation 163
5.3.3 Nonlinearity 163
5.3.4 Contact Resistance 166
5.4 Resistor Parasitics 167

5.5 Comparison of Available Resistors 170
5.5.1 Base Resistors 170
5.5.2 Emitter Resistors 171
5.5.3 Base Pinch Resistors 172
5.5.4 High-Sheet Resistors 173
5.5.5 Epi Pinch Resistors 175
5.5.6 Metal Resistors 176
5.5.7 Poly Resistors 177
5.5.8 NSD and PSD Resistors 180
5.5.9 N-well Resistors 180
5.5.10 Thin-film Resistors 181
5.6 Adjusting Resistor Values 182
5.6.1 Tweaking Resistors 182
Sliding Contacts 183
Sliding Heads 184
Trombone Slides 184
Metal Options 184
5.6.2 Trimming Resistors 185
Fuses 185
Zener Zaps 189
Laser Trims 190
5.7 Summary 191
5.8 Exercises 192

6 *Capacitors*
6.1 Capacitance 194
6.2 Capacitor Variability 200
6.2.1 Process Variation 200
6.2.2 Voltage Modulation and Temperature Variation 201
6.3 Capacitor Parasitics 203
6.4 Comparison of Available Capacitors 205
6.4.1 Base-emitter Junction Capacitors 205
6.4.2 MOS Capacitors 207
6.4.3 Poly-poly Capacitors 209
6.4.4 Miscellaneous Styles of Capacitors 211
6.5 Summary 212
6.6 Exercises 212

7 *Matching of Resistors and Capacitors*
7.1 Measuring Mismatch 214
7.2 Causes of Mismatch 217
7.2.1 Random Statistical Fluctuations 217
7.2.2 Process Biases 219
7.2.3 Pattern Shift 220

7.2.4 Variations in Polysilicon Etch Rate 222
7.2.5 Diffusion Interactions 224
7.2.6 Stress Gradients and Package Shifts 226
Piezoresistivity 227
Gradients and Centroids 229
Common-centroid Layout 231
Location and Orientation 235
7.2.7 Temperature Gradients and Thermoelectrics 236
Thermal Gradients 238
Thermoelectric Effects 240
7.2.8 Electrostatic Interactions 242
Voltage Modulation 242
Charge Spreading 245
Dielectric Polarization 246
Dielectric Relaxation 248
7.3 Rules for Device Matching 249
7.3.1 Rules for Resistor Matching 249
7.3.2 Rules for Capacitor Matching 253
7.4 Summary 257
7.5 Exercises 257

8 *Bipolar Transistors*
8.1 Topics in Bipolar Transistor Operation 260
8.1.1 Beta Rolloff 262
8.1.2 Avalanche Breakdown 262
8.1.3 Thermal Runaway and Secondary Breakdown 264
8.1.4 Saturation in NPN Transistors 266
8.1.5 Saturation in Lateral PNP Transistors 270
8.1.6 Parasitics of Bipolar Transistors 272
8.2 Standard Bipolar Small-signal Transistors 274
8.2.1 The Standard Bipolar NPN Transistor 274
Construction of Small-signal NPN Transistors 276
8.2.2 The Standard Bipolar Substrate PNP Transistor 279
Construction of Small-signal Substrate PNP Transistors 281
8.2.3 The Standard Bipolar Lateral PNP Transistor 283
Construction of Small-signal Lateral PNP Transistors 285
8.2.4 High-voltage Bipolar Transistors 291
8.3. Alternative Small-signal Bipolar Transistors 293
8.3.1 Extensions to Standard Bipolar 293
8.3.2 Analog BiCMOS Bipolar Transistors 294
8.3.3 Bipolar Transistors in a CMOS Process 297
8.3.4 Advanced-technology Bipolar Transistors 299
8.4 Summary 302
8.5 Exercises 303

9 *Applications of Bipolar Transistors*
9.1 Power Bipolar Transistors 306
9.1.1 Failure Mechanisms of NPN Power Transistors 307
Emitter Debiasing 307
Thermal Runaway and Secondary Breakdown 309
9.1.2 Layout of Power NPN Transistors 311
The Interdigitated-emitter Transistor 311
The Wide-emitter Narrow-contact Transistor 314
The Christmas-tree Device 315
The Cruciform-emitter Transistor 316
Power Transistor Layout in Analog BiCMOS 317
Selecting a Power Transistor Layout 318
9.1.3 Saturation Detection and Limiting 319
9.2 Matching Bipolar Transistors 322
9.2.1 Random Variations 323
9.2.2 Emitter Degeneration 325
9.2.3 NBL Shadow 327
9.2.4 Thermal Gradients 328
9.2.5 Stress Gradients 332
9.3 Rules for Bipolar Transistor Matching 334
9.3.1 Rules for Matching NPN Transistors 335
9.3.2 Rules for Matching Lateral PNP Transistors 337
9.4 Summary 340
9.5 Exercises 340

10 *Diodes*
10.1 Diodes in Standard Bipolar 343
10.1.1. Diode-connected Transistors 343
10.1.2 Zener Diodes 346
Surface Zener Diodes 347
Buried Zeners 349
10.1.3 Schottky Diodes 352
10.2 Diodes in CMOS and BiCMOS Processes 356
10.3 Matching Diodes 359
10.3.1 Matching PN Junction Diodes 359
10.3.2 Matching Zener Diodes 360
10.3.3 Matching Schottky Diodes 361
10.4 Summary 362
10.5 Exercises 362

11 *MOS Transistors*
11.1 Topics in MOS Transistor Operation 364
11.1.1 Modeling the MOS Transistor 364
Device Transconductance 365
Threshold Voltage 367

11.1.2 Parasitics of MOS Transistors 370
Breakdown Mechanisms 372
CMOS Latchup 375
11.2 Self-aligned Poly-Gate CMOS Transistors 376
11.2.1 Coding the MOS Transistor 377
Width and Length 378
11.2.2 N-well and P-well Processes 379
11.2.3 Channel Stops 381
11.2.4 Threshold Adjust Implants 383
11.2.5 Scaling the Transistor 386
11.2.6 Variant Structures 388
Serpentine Transistors 391
Annular Transistors 391
11.2.7 Backgate Contacts 393
11.3 Summary 396
11.4 Exercises 396

12 *Applications of MOS Transistors*
12.1 Extended-voltage Transistors 399
12.1.1 LDD and DDD Transistors 400
12.1.2 Extended-drain Transistors 403
Extended-drain NMOS Transistors 403
Extended-drain PMOS Transistors 405
12.1.3 Multiple Gate Oxides 405
12.2 Power MOS Transistors 407
Thermal Runaway 407
Secondary Breakdown 408
Rapid Transient Overload 408
MOS Switches versus Bipolar Switches 409
12.2.1 Conventional MOS Power Transistors 410
The Rectangular Device 411
The Diagonal Device 413
Computation of R_M 413
Other Considerations 414
Nonconventional Structures 416
12.2.2 DMOS Transistors 417
The Lateral DMOS Transistor 418
The DMOS NPN 420
12.3 The JFET Transistor 422
12.3.1 Modeling the JFET 422
12.3.2 JFET Layout 423
12.4 MOS Transistor Matching 426
12.4.1 Geometric Effects 427
Gate Area 428
Gate Oxide Thickness 428

Channel Length Modulation 429
Orientation 429
12.4.2 Diffusion and Etch Effects 430
Polysilicon Etch Rate Variations 430
Contacts Over Active Gate 431
Diffusions Near the Channel 432
PMOS versus NMOS Transistors 432
12.4.3 Thermal and Stress Effects 433
Oxide Thickness Gradients 433
Stress Gradients 433
Metallization-induced Stresses 434
Thermal Gradients 434
12.4.4 Common-centroid Layout of MOS Transistors 435
12.5 Rules for MOS Transistor Matching 439
12.6 Summary 442
12.7 Exercises 443

13 *Special Topics*
13.1 Merged Devices 445
13.1.1 Flawed Device Mergers 446
13.1.2 Successful Device Mergers 450
13.1.3 Low-risk Merged Devices 452
13.1.4 Medium-risk Merged Devices 453
13.1.5 Devising New Merged Devices 455
13.2 Guard Rings 455
13.2.1 Standard Bipolar Electron Guard Rings 456
13.2.2 Standard Bipolar Hole Guard Rings 457
13.2.3 Guard Rings in CMOS and BiCMOS Designs 458
13.3 Single-level Interconnection 460
13.3.1 Mock Layouts and Stick Diagrams 461
13.3.2 Techniques for Crossing Leads 463
13.3.3 Types of Tunnels 464
13.4 Constructing the Padring 466
13.4.1 Scribe Streets and Alignment Markers 466
13.4.2 Bondpads, Trimpads, and Testpads 468
13.4.3 ESD Structures 471
Zener Clamp 473
Two-stage Zener Clamps 475
Buffered Zener Clamp 476
V_{CES} Clamp 478
V_{CES} Clamp 479
Antiparallel Diode Clamps 480
Additional ESD Structures for CMOS Processes 480
13.4.4 Selecting ESD Structures 483
13.5 Exercises 485

14 *Assembling the Die*
14.1 Die Planning 488
14.1.1 Cell Area Estimation 489
Resistors 489
Capacitors 489
Vertical Bipolar Transistors 489
Lateral PNP Transistors 490
MOS Transistors 490
MOS Power Transistors 490
Computing Cell Area 491
14.1.2 Die Area Estimation 491
14.1.3 Gross Profit Margin 494
14.2 Floorplanning 495
14.3 Top-level Interconnection 500
14.3.1 Principles of Channel Routing 501
14.3.2 Special Routing Techniques 503
Kelvin Connections 503
Noisy Signals and Sensitive Signals 504
14.3.3 Electromigration 506
14.3.4 Minimizing Stress Effects 508
14.4 Conclusion 510
14.5 Exercises 510

Appendices
A. Table of Acronyms Used in the Text 513
B. The Miller Indices of a Cubic Crystal 516
C. Sample Layout Rules 519
D. Mathematical Derivations 527
E. Sources for Layout Editor Software 532

Index 533

Preface

An integrated circuit reveals its true appearance only under high magnification. The intricate tangle of microscopic wires covering its surface, and the equally intricate patterns of doped silicon beneath it, all follow a set of blueprints called a *layout*. The process of constructing layouts for analog and mixed-signal integrated circuits has stubbornly defied all attempts at automation. The shape and placement of every polygon require a thorough understanding of the principles of device physics, semiconductor fabrication, and circuit theory. Despite thirty years of research, much remains uncertain. What information there is lies buried in obscure journal articles and unpublished manuscripts. This textbook assembles this information between a single set of covers. While primarily intended for use by practicing layout designers, it should also prove valuable to circuit designers who desire a better understanding of the relationship between circuits and layouts.

The text has been written for a broad audience, some of whom have had only limited exposure to higher mathematics and solid-state physics. The amount of mathematics has been kept to an absolute minimum, and care has been taken to identify all variables and to use the most accessible units. The reader need only have a familiarity with basic algebra and elementary electronics. Many of the exercises assume that the reader also has access to layout editing software, but those who lack such resources can complete many of the exercises using pencil and paper.

The text consists of fourteen chapters and five appendices. The first two chapters provide an overview of device physics and semiconductor processing. These chapters avoid mathematical derivations and instead emphasize simple verbal explanations and visual models. The third chapter presents three archetypal processes: standard bipolar, silicon-gate CMOS, and analog BiCMOS. The presentation focuses upon development of cross sections and the correlation of these cross sections to conventional layout views of sample devices. The fourth chapter covers common failure mechanisms and emphasizes the role of layout in determining reliability. Chapters Five and Six cover the layout of resistors and capacitors. Chapter Seven presents the principles of matching, using resistors and capacitors as examples. Chapters Eight through Ten cover the layout of bipolar devices, while chapters Eleven and Twelve cover the layout and matching of field-effect transistors. Chapters Thirteen and Fourteen cover a variety of advanced topics, including device mergers, guard rings, ESD protection structures, and floorplanning. The appendices include a list of acronyms, a discussion of Miller indices, sample layout rules for use in working the exercises, and the derivation of formulas used in the text.

Alan Hastings

Acknowledgments

The information contained in this text has been gathered through the hard work of many scientists, engineers, and technicians, the vast majority of whom must remain unacknowledged because their work has not been published. I have included references to as many fundamental discoveries and principles as I could, but in many cases I have been unable to determine original sources.

I thank my colleagues at Texas Instruments for numerous suggestions. I am especially grateful to Ken Bell, Walter Bucksch, Lou Hutter, Clif Jones, Jeff Smith, Fred Trafton, and Joe Trogolo, all of whom have provided important information for this text. I am also grateful for the encouragement of Bob Borden, Nicolas Salamina, and Ming Chiang, without which this text would never have been written.

1 Device Physics

Before 1960, most electronic circuits depended upon vacuum tubes to perform the critical tasks of amplification and rectification. An ordinary mass-produced AM radio required five tubes, while a color television needed no fewer than twenty. Vacuum tubes were large, fragile, and expensive. They dissipated a lot of heat and were not very reliable. So long as electronics depended upon them, it was nearly impossible to construct systems requiring thousands or millions of active devices.

The appearance of the bipolar junction transistor in 1947 marked the beginning of the solid-state revolution. These new devices were small, cheap, rugged, and reliable. Solid-state circuitry made possible the development of pocket transistor radios and hearing aids, quartz watches and touch-tone phones, compact disc players and personal computers.

A *solid-state device* consists of a crystal with regions of impurities incorporated into its surface. These impurities modify the electrical properties of the crystal, allowing it to amplify or modulate electrical signals. A working knowledge of device physics is necessary to understand how this occurs. This chapter covers not only elementary device physics but also the operation of three of the most important solid-state devices: the junction diode, the bipolar transistor, and the field-effect transistor. Chapter 2 explains the manufacturing processes used to construct these and other solid-state devices.

1.1 SEMICONDUCTORS

The inside front cover of the book depicts a long-form periodic table. The elements are arranged so those with similar properties group together to form rows and columns. The elements on the left-hand side of the periodic table are called *metals*, while those on the right-hand side are called *nonmetals*. Metals are usually good conductors of heat and electricity. They are also malleable and display a characteristic metallic luster. Nonmetals are poor conductors of heat and electricity, and those that are solid are brittle and lack the shiny luster of metals. A few elements in the middle

of the periodic table, such as silicon and germanium, have electrical properties that lie midway between those of metals and nonmetals. These elements are called *semiconductors*. The differences between metals, semiconductors, and nonmetals result from the electronic structure of their respective atoms.

Every atom consists of a positively charged nucleus surrounded by a cloud of electrons. The number of electrons in this cloud equals the number of protons in the nucleus, which also equals the atomic number of the element. Therefore a carbon atom has six electrons because carbon has an atomic number of six. These electrons occupy a series of *shells* that are somewhat analogous to the layers of an onion. As electrons are added, the shells fill in order from innermost outward. The outermost or *valence* shell may remain unfilled. The electrons occupying this outermost shell are called *valence electrons*. The number of valence electrons possessed by an element determines most of its chemical and electronic properties.

Each row of the periodic table corresponds to the filling of one shell. The leftmost element in the row has one valence electron, while the rightmost element has a full valence shell. Atoms with filled valence shells possess a particularly favored configuration. Those with unfilled valence shells will trade or share electrons so that each can claim a full shell. Electrostatic attraction forms a chemical bond between atoms that trade or share electrons. Depending upon the strategy adopted to fill the valence shell, one of three types of bonding will occur.

Metallic bonding occurs between atoms of metallic elements, such as sodium. Consider a group of sodium atoms in close proximity. Each atom has one valence electron orbiting around a filled inner shell. Imagine that the sodium atoms all discard their valence electrons. The discarded electrons are still attracted to the positively charged sodium atoms, but, since each atom now has a full valence shell, none accepts them. Figure 1.1A shows a simplified representation of a sodium crystal. Electrostatic forces hold the sodium atoms in a regular lattice. The discarded valence electrons wander freely through the resulting crystal. Sodium metal is an excellent electrical conductor due to the presence of numerous free electrons.[1] These same electrons are also responsible for the metallic luster of the element and its high thermal conductivity. Other metals form similar crystal structures, all of which are held together by metallic bonding between a sea of free valence electrons and a rigid lattice of charged atomic cores.

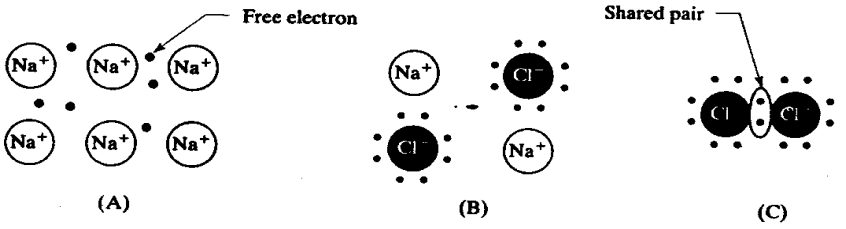


FIGURE 1.1 Simplified illustrations of various types of chemical bonding: metallically bonded sodium crystal (A), ionically bonded sodium chloride crystal (B), and covalently bonded chlorine molecule (C).

Ionic bonding occurs between atoms of metals and nonmetals. Consider a sodium atom in close proximity to a chlorine atom. The sodium atom has one valence electron, while the chlorine atom is one electron short of a full valence shell. The sodium atom can donate an electron to the chlorine atom and by this means both can achieve filled outer shells. After the exchange, the sodium atom has a net posi-

[1] Some metals conduct by means of holes rather than electrons, but the general observations made in the text still apply.

tive charge and the chlorine atom a net negative charge. The two charged atoms (or *ions*) attract one another. Solid sodium chloride thus consists of sodium and chlorine ions arranged in a regular lattice, forming a crystal (Figure 1.1B). Crystalline sodium chloride is a poor conductor of electricity, since all of its electrons are held in the shells of the various atoms.

Covalent bonding occurs between atoms of nonmetals. Consider two chlorine atoms in close proximity. Each atom has only seven valence electrons, while each needs eight to fill its valence shell. Suppose that each of the two atoms contributes one valence electron to a common pair shared by both. Now each chlorine atom can claim eight valence electrons: six of its own, plus the two shared electrons. The two chlorine atoms link to form a molecule that is held together by the electron pair shared between them (Figure 1.1C). The shared pair of electrons forms a *covalent bond.* The lack of free valence electrons explains why nonmetallic elements do not conduct electricity and why they lack metallic luster. Many nonmetals are gases at room temperature because the electrically neutral molecules exhibit no strong attraction to one another and thus do not condense to form a liquid or a solid.

The atoms of a semiconductor also form covalent bonds. Consider atoms of silicon, a representative semiconductor. Each atom has four valence electrons and needs four more to complete its valence shell. Two silicon atoms could theoretically attempt to pool their valence electrons to achieve filled shells. In practice this does not occur because eight electrons packed tightly together strongly repel one another. Instead, each silicon atom shares one electron pair with each of four surrounding atoms. In this way, the valence electrons are spread around to four separate locations and their mutual repulsion is minimized.

Figure 1.2 shows a simplified representation of a silicon crystal. Each of the small circles represents a silicon atom. Each of the lines between the circles represents a covalent bond consisting of a shared pair of valence electrons. Each silicon atom can claim eight electrons (four shared electron pairs), so all of the atoms have full valence shells. These atoms are linked together in a molecular network by the covalent bonds formed between them. This infinite lattice represents the structure of the silicon crystal. The entire crystal is literally a single molecule, so crystalline silicon is strong and hard, and it melts at a very high temperature. Silicon is normally a poor conductor of electricity because all of its valence electrons are used to form the crystal lattice.

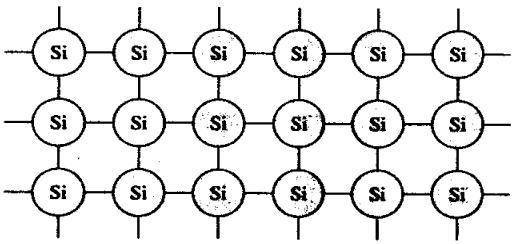


FIGURE 1.2 Simplified two-dimensional representation of a silicon crystal lattice.

A similar macromolecular crystal can theoretically be formed by any group-IV element,[2] including carbon, silicon, germanium, tin, and lead. Carbon, in the form of diamond, has the strongest bonds of any group-IV element. Diamond crystals

[2] The group-III, IV, V, and VI elements reside in columns III-B, IV-B, V-B, and VI-B of the long-form periodic table. The group-II elements may fall into either columns II-A or II-B. The A/B numbering system is a historical curiosity and the International Union of Pure and Applied Chemists (IUPAC) has recommended its abandonment; see J. Hudson, *The History of Chemistry* (New York: Chapman and Hall, 1992), pp. 122–137.

are justly famed for their strength and hardness. Silicon and germanium have somewhat weaker bonds due to the presence of filled inner shells that partially shield the valence electrons from the nucleus. Tin and lead have weak bonds because of numerous inner shells; they typically form metallically bonded crystals instead of covalently bonded macromolecules. Of the group-IV elements, only silicon and germanium have bonds of an intermediate degree of strength. These two act as true semiconductors, while carbon is a nonmetal, and tin and lead are both metals.

1.1.1. Generation and Recombination

The electrical conductivity of group-IV elements increases with atomic number. Carbon, in the form of diamond, is a true insulator. Silicon and germanium have much higher conductivities, but these are still far less than those of metals such as tin and lead. Because of their intermediate conductivities, silicon and germanium are termed *semiconductors*.

Conduction implies the presence of free electrons. At least a few of the valence electrons of a semiconductor must somehow escape the lattice to support conduction. Experiments do indeed detect small but measurable concentrations of free electrons in pure silicon and germanium. The presence of these free electrons implies that some mechanism provides the energy needed to break the covalent bonds. The statistical theory of thermodynamics suggests that the source of this energy lies in the random thermal vibrations that agitate the crystal lattice. Even though the average thermal energy of an electron is relatively small (less than 0.1 electron volt), these energies are randomly distributed, and a few electrons possess much larger energies. The energy required to free a valence electron from the crystal lattice is called the *bandgap energy*. A material with a large bandgap energy possesses strong covalent bonds and therefore contains few free electrons. Materials with lower bandgap energies contain more free electrons and possess correspondingly greater conductivities (Table 1.1).

TABLE 1.1 Selected properties of group-IV elements.[3]

Element	Atomic Number	Melting Point, °C	Electrical Conductivity $(\Omega cm)^{-1}$	Bandgap Energy, eV
Carbon (diamond)	6	3550	$\sim 10^{-16}$	5.2
Silicon	14	1410	$4 \cdot 10^{-6}$	1.1
Germanium	32	937	0.02	0.7
White Tin	50	232	$9 \cdot 10^{4}$	0.1

A vacancy occurs whenever an electron leaves the lattice. One of the atoms that formerly possessed a full outer shell now lacks a valence electron and therefore has a net positive charge. This situation is depicted in a simplified fashion in Figure 1.3. The ionized atom can regain a full valence shell if it appropriates an electron from a neighboring atom. This is easily accomplished since it still shares electrons with three adjacent atoms. The electron vacancy is not eliminated; it merely shifts to the

[3] Bandgap energies for Si, Ge: B. G. Streetman, *Solid State Electronic Devices*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1980), p. 443. Bandgap for C: N. B. Hanny, ed., *Semiconductors* (New York: Reinhold Publishing, 1959), p. 52. Conductivity for Sn: R. C. Weast, ed., *CRC Handbook of Chemistry and Physics*, 62nd ed. (Boca Raton, FL: CRC Press, 1981), pp. F135–F136. Other values computed. Melting points: Weast, pp. B4–B48.

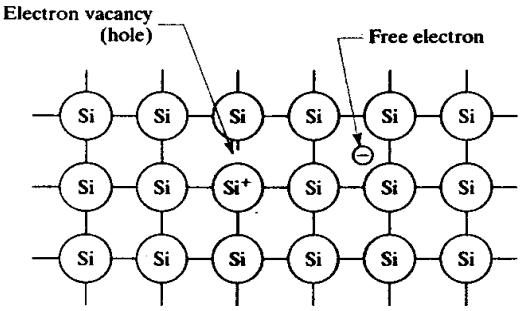


FIGURE 1.3 Simplified diagram of thermal generation in intrinsic silicon.

adjacent atom. As the vacancy is handed from atom to atom, it moves through the lattice. This moving electron vacancy is called a *hole.*

Suppose an electric field is placed across the crystal. The negatively charged free electrons move toward the positive end of the crystal. The holes behave as if they were positively charged particles and move toward the negative end of the crystal. The motion of the holes can be compared to bubbles in a liquid. Just as a bubble is a location devoid of fluid, a hole is a location devoid of valence electrons. Bubbles move upward because the fluid around them sinks downward. Holes shift toward the negative end of the crystal because the surrounding electrons shift toward the positive end.

Holes are usually treated as if they were actual subatomic particles. The movement of a hole toward the negative end of the crystal is explained by assuming that holes are positively charged. Similarly, their rate of movement through the crystal is measured by a quantity called *mobility.* Holes have lower mobilities than electrons; typical values in bulk silicon are $480 cm^2/V \cdot sec$ for holes and $1350 cm^2/V \cdot sec$ for electrons.[4] The lower mobility of holes makes them less efficient charge carriers. The behavior of a device therefore relies upon whether its operation involves holes or electrons.

A free electron and a hole are formed whenever a valence electron is removed from the lattice. Both particles are electrically charged and move under the influence of electric fields. Electrons move toward positive potentials, producing an electron current. Holes move toward negative potentials, producing a hole current. The total current equals the sum of the electron and the hole currents. Holes and electrons are both called *carriers* because of their role in transporting electric charge.

Carriers are always generated in pairs since the removal of a valence electron from the lattice simultaneously forms a hole. The generation of electron-hole pairs can occur whenever energy is absorbed by the lattice. Thermal vibration produces carriers, as do light, nuclear radiation, electron bombardment, rapid heating, mechanical friction, and any number of other processes. To consider only one example, light of a sufficiently short wavelength can generate electron-hole pairs. When a lattice atom absorbs a photon, the resulting energy transfer can break a covalent bond to produce a free electron and a free hole. Optical generation will occur only if the photons have enough energy to break bonds, and this in turn requires light of a sufficiently short wavelength. Visible light has enough energy to produce electron-hole pairs in most semiconductors. Solar cells make use of this phenomenon to convert sunlight into electrical current. Photocells and solid-state camera detectors also employ optical generation.

[4] Streetman, p. 443.

Just as carriers are generated in pairs, they also recombine in pairs. The exact mechanism of carrier recombination depends on the nature of the semiconductor. Recombination is particularly simple in the case of a *direct-bandgap semiconductor.* When an electron and a hole collide, the electron falls into the hole and repairs the broken covalent bond. The energy gained by the electron is radiated away as a photon (Figure 1.4A). Direct-bandgap semiconductors can, when properly stimulated, emit light. A *light-emitting diode* (LED) produces light by electron-hole recombination. The color of light emitted by the LED depends on the bandgap energy of the semiconductor used to manufacture it. Similarly, the so-called *phosphors* used in manufacturing glow-in-the-dark paints and plastics also contain direct-bandgap semiconductors. Electron-hole pairs form whenever the phosphor is exposed to light. A large number of electrons and holes gradually accumulate in the phosphor. The slow recombination of these carriers causes the emission of light.

Silicon and germanium are *indirect-bandgap semiconductors.* In these semiconductors, the collision of a hole and an electron will not cause the two carriers to recombine. The electron may momentarily fall into the hole, but quantum mechanical considerations prevent the generation of a photon. Since the electron cannot shed excess energy, it is quickly ejected from the lattice and the electron-hole pair reforms. In the case of an indirect-bandgap semiconductor, recombination can only occur at specific sites in the lattice, called *traps,* where flaws or foreign atoms distort the lattice (Figure 1.4B). A trap can momentarily capture a passing carrier. The trapped carrier becomes vulnerable to recombination because the trap can absorb the liberated energy.

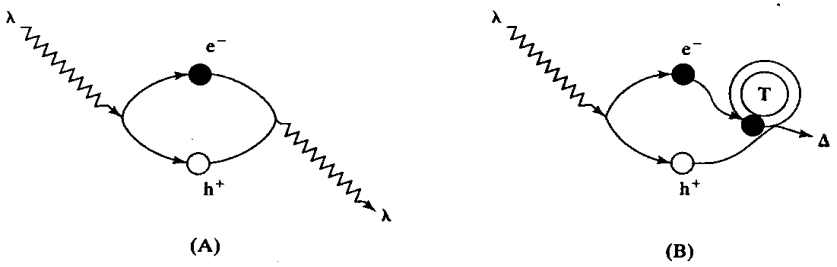


FIGURE 1.4 Schematic representations of recombination processes: (A) direct recombination, in which a photon, λ, generates a hole, h^+, and an electron, e^-, that collide and re-emit a photon; and (B) indirect recombination, in which one of the carriers is caught by a trap, T, and recombination takes place at the trap site with the liberation of heat, Δ.

Traps that aid the recombination of carriers are called *recombination centers.* The more recombination centers a semiconductor contains, the shorter the average time between the generation of a carrier and its recombination. This quantity, called the *carrier lifetime,* limits how rapidly a semiconductor device can switch on and off. Recombination centers are sometimes deliberately added to semiconductors to increase switching speeds. Gold atoms form highly efficient recombination centers in silicon, so high-speed diodes and transistors are sometimes made from silicon containing a small amount of gold. Gold is not the only substance that can form recombination centers. Many transition metals such as iron and nickel have a similar (if less potent) effect. Some types of crystal defects can also serve as recombination centers. Solid-state devices must be fabricated from extremely pure single-crystal materials in order to ensure consistent electrical performance.

1.1.2. Extrinsic Semiconductors

The conductivity of semiconductors depends upon their purity. Absolutely pure, or *intrinsic,* semiconductors have low conductivities because they contain only a few thermally generated carriers. The addition of certain impurities greatly increases the

number of available carriers. These *doped,* or *extrinsic,* semiconductors can approach the conductivity of a metal. A lightly doped semiconductor may contain only a few parts per billion of dopant. Even a heavily doped semiconductor contains only a few hundred parts per million due to the limited solid solubility of dopants in silicon. The extreme sensitivity of semiconductors to the presence of dopants makes it nearly impossible to manufacture truly intrinsic material. Practical semiconductor devices are, therefore, fabricated almost exclusively from extrinsic material.

Phosphorus-doped silicon is an example of an extrinsic semiconductor. Suppose a small quantity of phosphorus is added to a silicon crystal. The phosphorus atoms are incorporated into the crystal lattice in positions that would otherwise have been occupied by silicon atoms (Figure 1.5). Phosphorus, a group-V element, has five valence electrons. The phosphorus atom shares four of these with its four neighboring atoms. Four bonding electron pairs give the phosphorus atom a total of eight shared electrons. These, combined with the one remaining unshared electron, result in a total of nine valence electrons. Since eight electrons entirely fill the valence shell, no room remains for the ninth electron. This electron is expelled from the phosphorus atom and wanders freely through the crystal lattice. Each phosphorus atom added to the silicon lattice thus generates one free electron.

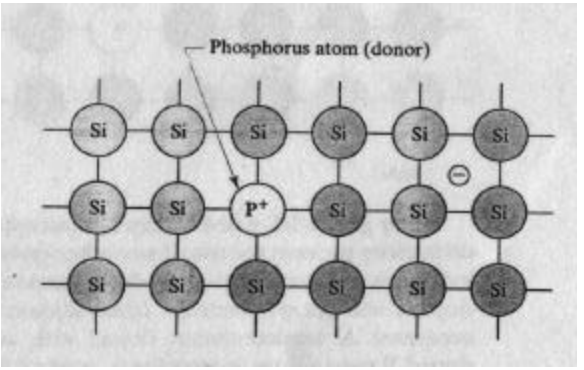


FIGURE 1.5 Simplified crystal structure of phosphorus-doped silicon.

The loss of the ninth electron leaves the phosphorus atom with a net positive charge. Although this atom is ionized, it does not constitute a hole. Holes are electron vacancies created by the removal of electrons from a filled valence shell. The phosphorus atom has a full valence shell despite its positive charge. The charge associated with the ionized phosphorus atom is therefore immobile.

Other group-V elements will have the same effect as phosphorus. Each atom of a group-V element that is added to the lattice will produce one additional free electron. Elements that donate electrons to a semiconductor in this manner are called *donors.* Arsenic, antimony, and phosphorus are all used in semiconductor processing as donors for silicon.

A semiconductor doped with a large number of donors has a preponderance of electrons as carriers. A few thermally generated holes still exist, but their numbers actually diminish in the presence of extra electrons. This occurs because the extra electrons increase the probability that the hole will find an electron and recombine. The large number of free electrons in N-type silicon greatly increases its conductivity (and greatly reduces its resistance).

A semiconductor doped with donors is said to be *N-type.* Heavily doped N-type silicon is sometimes marked N+, and lightly doped N-type silicon N–. The plus and minus symbols denote the relative numbers of donors, not electrical charges. Electrons are considered the *majority carriers* in N-type silicon due to their large

numbers. Similarly, holes are considered the *minority carriers* in N-type silicon. Strictly speaking, intrinsic silicon has neither majority nor minority carriers because both types are present in equal numbers.

Boron-doped silicon forms another type of extrinsic semiconductor. Suppose a small number of boron atoms are added to the silicon lattice (Figure 1.6). Boron, a group-III element, has three valence electrons. The boron atom attempts to share its valence electrons with its four neighboring atoms, but, because it has only three, it cannot complete the fourth bond. As a result, there are only seven valence electrons around the boron atom. The electron vacancy thus formed constitutes a hole. This hole is mobile and soon moves away from the boron atom. Once the hole departs, the boron atom is left with a negative charge caused by the presence of an extra electron in its valence shell. As in the case of phosphorus, this charge is immobile and does not contribute to conduction. Each atom of boron added to the silicon contributes one mobile hole.

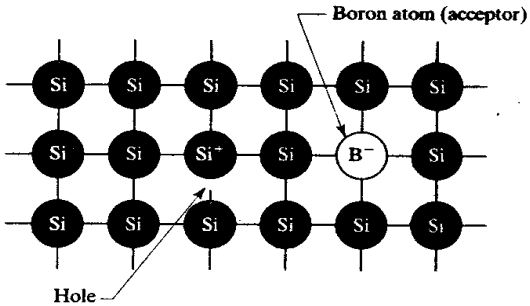


FIGURE 1.6 Simplified crystal structure of boron-doped silicon.

Other group-III elements can also accept electrons and generate holes. Technical difficulties prevent the use of any other group-III elements in silicon fabrication, but indium is sometimes used to dope germanium. Any group-III element used as a dopant will *accept* electrons from adjoining atoms, so these elements are called *acceptors*. A semiconductor doped with acceptors is said to be *P-type*. Heavily doped P-type silicon is sometimes marked P+, and lightly doped P-type silicon P–. Holes are the majority carriers and electrons are the minority carriers in P-type silicon. Table 1.2 summarizes some of the terminology used to describe extrinsic semiconductors.

TABLE 1.2 Extrinsic semiconductor terminology.

Semiconductor Type	Dopant Type	Typical Dopants for Silicon	Majority Carriers	Minority Carriers
N-type	Donors	Phosphorus, arsenic, and antimony	Electrons	Holes
P-type	Acceptors	Boron	Holes	Electrons

A semiconductor can be doped with both acceptors and donors. The dopant present in excess determines the type of the silicon and the concentration of the carriers. It is thus possible to invert P-type silicon to N-type by adding an excess of donors. Similarly, it is possible to invert N-type silicon to P-type by adding an excess of acceptors. The deliberate addition of an opposite-polarity dopant to invert the type of a semiconductor is called *counterdoping*. Most modern semiconductors are

made by selectively counterdoping silicon to form a series of P- and N-type regions. Much more will be said about this practice in the next chapter.

If counterdoping were taken to extremes, the entire crystal lattice would consist of an equal ratio of acceptor and donor atoms. The two types of atoms would be present in exactly equal numbers. The resulting crystal would have very few free carriers and would appear to be an intrinsic semiconductor. Such *compound semiconductors* actually exist. The most familiar example is *gallium arsenide*, a compound of gallium (a group-III element) and arsenic (a group-V element). Materials of this sort are called III-V compound semiconductors. They include not only gallium arsenide but also gallium phosphide, indium antimonide, and many others. Many III-V compounds are direct-bandgap semiconductors, and some are used in constructing light-emitting diodes and semiconductor lasers. Gallium arsenide is also employed to a limited extent for manufacturing very high-speed solid-state devices, including integrated circuits. II-VI compound semiconductors are composed of equal mixtures of group-II and group-VI elements. Cadmium sulfide is a typical II-VI compound used to construct photosensors. Other II-VI compounds are used as phosphors in cathode ray tubes. A final class of semiconductors includes IV-IV compounds such as silicon carbide, recently used on a small scale to fabricate blue LEDs.

Of all the semiconductors, only silicon possesses the physical properties required for high-volume, low-cost manufacture of integrated circuits. The vast majority of solid-state devices are fabricated in silicon, and all other semiconductors are relegated to niche markets. The remainder of this text, therefore, focuses upon silicon integrated circuits.

1.1.3. Diffusion and Drift

The motion of carriers through a silicon crystal results from two separate processes: *diffusion* and *drift*. Diffusion is a random motion of carriers that occurs at all times and places, while drift is a unidirectional movement of carriers under the influence of an electric field. Both of these processes contribute to conduction in semiconductors.

Diffusion closely resembles Brownian motion. That is, individual carriers move through the semiconductor until they collide with lattice atoms. The collision process scatters the carriers through unpredictable angles. After a very few collisions, the motion of the carriers becomes completely randomized. The carriers wander aimlessly about, tracing a sort of drunkard's walk (Figure 1.7A).

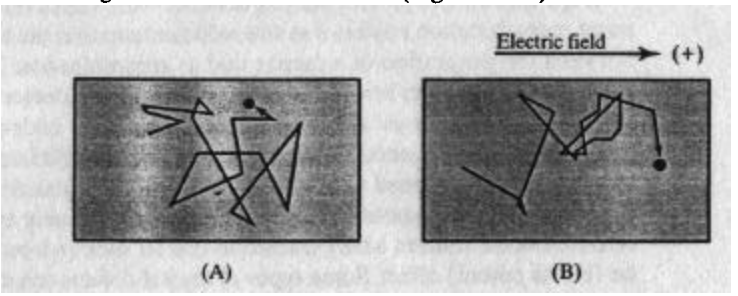


FIGURE 1.7 Comparison of conduction mechanisms for an electron: diffusion (A) and drift superimposed on diffusion (B). Notice the gradual motion of the electron toward the positive potential.

The diffusion of carriers through a semiconductor is analogous to the diffusion of dye molecules through still water. When a drop of concentrated dye falls into water, the dye molecules all initially occupy a small volume of liquid. The molecules gradually diffuse from regions of higher concentration to regions of lower concentration. Eventually the dye becomes distributed uniformly throughout the solution. Similarly, the diffusion of carriers across concentration gradients produces a *diffusion current*.

Unless some mechanism constantly adds more carriers, diffusion eventually redistributes them uniformly throughout the silicon and the diffusion current subsides.

The motion of a carrier under the influence of an electric field is called *drift*. Although the carrier still collides with the lattice and thus moves in a random drunkard's walk, it gradually drifts in a specific direction (Figure 1.7B). This subtle bias is caused by the electric field. No matter what direction the carrier moves, the field always and relentlessly acts upon it. If the carrier is moving opposite the field, its motion is retarded; if it is moving with the field, its motion is accelerated. Frequent collisions prevent the carrier from building up any appreciable velocity, but a subtle overall motion appears. Electrons move toward positive potentials, even if slowly and erratically. Holes likewise move toward negative potentials. A simple analogy to drift consists of the motion of the steel ball in a pinball machine. Although bumpers and pegs may divert the ball in any direction, the tilt of the board causes it to eventually move downward. Similarly, an electric field biases carriers toward motion in a specific direction, producing a *drift current*.

1.2 PN JUNCTIONS

Uniformly doped semiconductors have few applications. Almost all solid-state devices contain a combination of multiple P- and N-type regions. The interface between a P-type region and an N-type region is called a *PN junction*, or simply a *junction*.

Figure 1.8A shows two pieces of silicon. On the left is a bar of P-type silicon, and on the right is a bar of N-type silicon. No junction is present as long as the two are not in contact with one another. Each piece of silicon contains a uniform distribution of carriers. The P-type silicon has a large majority of holes and a few electrons; the N-type silicon has a large majority of electrons and a few holes.

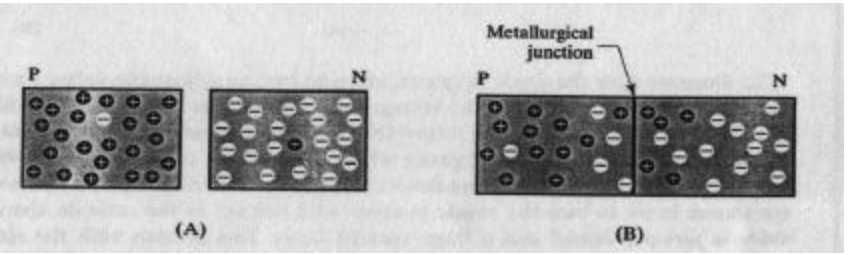


FIGURE 1.8 Carrier populations in silicon before the junction is assembled (A), and afterward (B).

Now, imagine the two pieces of silicon are brought into contact with one another to form a junction. No physical barrier to the motion of the carriers remains. There is a great excess of holes in the P-type silicon, and a great excess of electrons in the N-type silicon. Some of the holes diffuse from the P-type silicon to the N-type. Likewise, some of the electrons diffuse from the N-type silicon to the P-type. Figure 1.8B shows the result. A number of carriers have diffused across the junction in both directions. The concentration of minority carriers on either side of the junction has risen above that which would be produced by doping alone. The excess of minority carriers produced by diffusion across a junction is called the *excess minority carrier concentration*.

1.2.1. Depletion Regions

The presence of excess minority carriers on either side of a junction has two effects. First, the carriers produce an electric field. The extra holes in the N-type silicon represent a positive charge, while the excess electrons in the P-type silicon represent a

negative charge. Thus an electric potential develops across the PN junction that biases the N-side of the junction positive with respect to the P-side.

When carriers diffuse across the junction, they leave equal numbers of ionized dopant atoms behind. These atoms are rigidly fixed in the crystal lattice and cannot move. On the P-side of the junction lie ionized acceptors that produce a negative charge. On the N-side of the junction lie ionized donors that produce a positive charge. An electric potential again develops that biases the N-side of the junction positive with respect to the P-side. This potential adds to the one produced by the separation of the charged carriers.

Carriers tend to drift in the presence of an electric field. Holes are attracted to the negative potential on the P-side of the junction. Similarly, electrons are attracted to the positive potential on the N-side of the junction. The drift of carriers thus tends to oppose their diffusion. Holes diffuse from the P-side of the junction to the N-side and drift back. Electrons diffuse from the N-side of the junction to the P-side and drift back. Equilibrium occurs when the drift and diffusion currents are equal and opposite. The excess minority carrier concentrations on either side of the junction also reach equilibrium values, as does the voltage potential across the junction.

The voltage difference across a PN junction in equilibrium is called its *built-in potential,* or its *contact potential.* In a typical silicon PN junction, the built-in potential can range from a few tenths of a volt to as much as a volt. Heavily doped junctions have larger built-in potentials than lightly doped ones. Because of the higher doping levels, more carriers diffuse across the heavily doped junction and thus a larger diffusion current flows. In order to restore equilibrium, a larger drift current is also needed and thus a stronger electric field develops. Heavily doped junctions therefore have larger built-in potentials than lightly doped ones.

Although the built-in potential is quite real, it cannot be measured with a voltmeter. This apparent paradox can be explained by a closer examination of a circuit containing a PN junction and a voltmeter (Figure 1.9). The two probes of the meter are made of metal, not silicon. The points of contact between the metal probes and the silicon also form junctions, each of which has a contact potential of its own. Because the silicon beneath the two probes has different doping levels, the two contact potentials of the probe points are unequal. The difference between these two contact potentials exactly cancels the built-in potential of the PN junction, and no current flows in the external circuit. This situation must occur because any current flow would constitute a free energy source, or a sort of perpetual motion machine. The cancellation of the built-in potentials ensures that energy cannot be extracted from a PN junction in equilibrium and thus prevents a violation of the laws of thermodynamics.

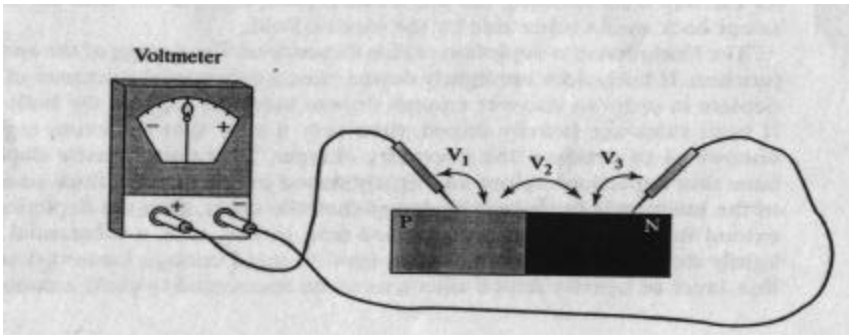


FIGURE 1.9 Demonstration of the impossibility of directly measuring a built-in potential. Contact potentials V_1 and V_3 exactly cancel built-in potential V_2.

The built-in potential has two causes: the separation of ionized dopant atoms and the separation of charged carriers. The carriers are free to move, but the dopant atoms are rigidly fixed in the crystal lattice. If the dopant atoms could move, they would be drawn together by their opposite charges. They remain separated because they are anchored to the lattice. The region occupied by these charged atoms is subject to a strong electric field. Any carrier that enters this region must move quickly or it will be swept out again by the field. As a result, this region contains very few carriers at any given instant in time. This region is sometimes called a *space charge layer* because of the presence of the charged dopant atoms. More commonly, it is called a *depletion region* because of the relatively low concentration of carriers found there.

If the depletion region contains few carriers, then the excess minority carriers must pile up on either side of it. Figure 1.10 graphically shows the resulting distributions of excess minority carriers. The concentration gradients cause these carriers to diffuse into the electrically neutral regions beyond the junction. The electric field produced by the separation of these charged carriers pulls them back toward the junction. An equilibrium is soon established, resulting in steady-state distributions of minority carriers resembling those shown in Figure 1.10.

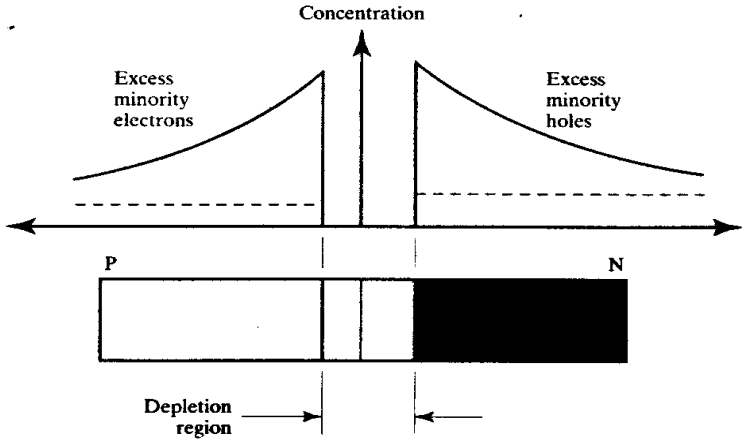


FIGURE 1.10 Diagram of excess minority carrier concentrations on either side of a PN junction in equilibrium.

The behavior of a PN junction can be summarized as follows: the diffusion of carriers across the junction produces excess minority carrier concentrations on either side of a depletion region. The separation of ionized dopant atoms causes an electric field to form across the depletion region. This field prevents most of the majority carriers from crossing the depletion region, and the few that do are eventually swept back to the other side by the electric field.

The thickness of a depletion region depends on the doping of the two sides of the junction. If both sides are lightly doped, then a substantial thickness of silicon must deplete in order to uncover enough dopant atoms to support the built-in potential. If both sides are heavily doped, then only a very thin depletion region need be uncovered to produce the necessary charges. Therefore heavily doped junctions have thin depletion regions and lightly doped junctions have thick ones. If one side of the junction is more heavily doped than the other, then the depletion region will extend further into the lightly doped side. In this case, a substantial thickness of lightly doped silicon must be uncovered to yield enough ionized dopants. Only a thin layer of heavily doped silicon need be uncovered to yield a counterbalancing

charge. Figure 1.10 illustrates this case since the N-side of the junction is more lightly doped than the P-side.

1.2.2. PN Diodes

A PN junction forms a very useful solid-state device called a *diode.* Figure 1.11 shows a simplified diagram of the structure of a PN diode. The diode has, as its name suggests, two terminals. One terminal, called the *anode,* connects to the P-side of the junction. The other terminal, called the *cathode,* connects to the N-side. These two terminals are used to connect the diode to an electrical circuit. The schematic symbol for a diode consists of an arrowhead representing the anode and a perpendicular line representing the cathode. Diodes conduct current preferentially in one direction—that indicated by the arrowhead.

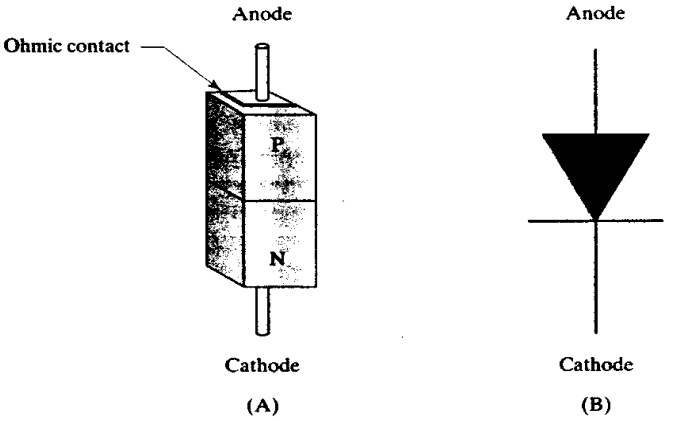


FIGURE 1.11 PN junction diode: simplified structure (A) and standard schematic symbol (B).

To illustrate how the diode operates, imagine that an adjustable voltage source has been connected across it. If the voltage source is set to zero volts, then the diode is under *zero bias.* No current will flow through a zero-biased diode. If the voltage source is set to bias the anode negative with respect to the cathode, then the diode is *reverse biased.* Very little current flows through a reverse-biased diode. If the voltage source is set to bias the anode positive with respect to the cathode, then the diode is *forward biased* and a large current flows. This accords with the simple mnemonic: *current flows with the arrow, not against it.* Devices that conduct in only one direction are called *rectifiers.* They find frequent application in power supplies, radio receivers, and signal processing circuits.

Diode rectification depends upon the presence of a junction. Each of the three bias conditions can be explained by an appropriate analysis of carrier flows across this junction. The case of the zero-bias diode is particularly simple since it is identical to the case of the equilibrium junction already discussed. The only potential present across the junction is the built-in potential. When the diode is connected in a circuit, the contact potentials of the leads touching the silicon balance the built-in potential of the junction. Thus no current flows in the circuit.

The behavior of a reverse-biased diode is also simple to explain. The reverse bias makes the N-side of the junction even more positive with respect to the P-side. The voltage seen across the junction increases, so excess minority carriers continue to be swept back across it and majority carriers continue to be held on their respective sides of the junction. The increased voltage across the junction causes the ionization

of additional dopant atoms on either side, so the depletion region widens as the reverse bias increases.

The behavior of a forward-biased junction is somewhat more complex. The voltage applied to the terminals opposes the built-in potential. The voltage across the junction therefore lessens and the depletion region thins. The drift currents caused by the electric field are simultaneously reduced. More and more majority carriers make the transit across the depletion region without being swept back by the electric field. Figure 1.12 shows graphically the overall flow of carriers: holes are injected across the junction from anode to cathode (left to right), while electrons are injected across the junction from cathode to anode (right to left). In the illustrated diode, the hole current across the junction outweighs the electron current because the anode is more heavily doped than the cathode and there are more majority holes available in the anode than there are majority electrons in the cathode. Once these carriers have been injected across the junction, they become minority carriers and recombine with majority carriers present on the other side. Currents are drawn in from the terminals in order to replenish the supply of majority carriers in the neutral silicon. This illustration is somewhat simplified since it only shows the general flow of carriers through the diode. Some of the carriers injected across the junction are swept back by the electric field before they can recombine. Such carriers do not contribute to the net current flow through the diode, so they are not illustrated. Likewise, the tiny numbers of thermally generated minority carriers that cross the junction are not shown since they form an insignificant portion of the overall current flow through a forward-biased diode.

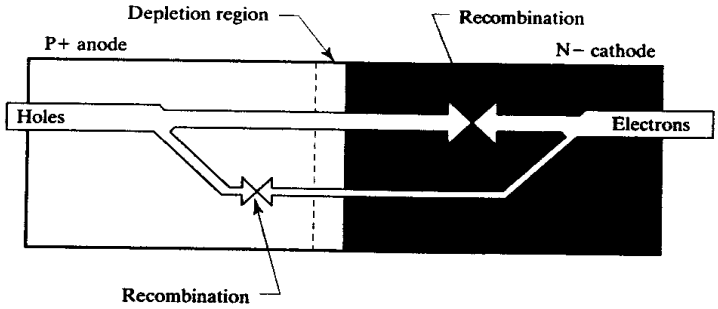


FIGURE 1.12 Carrier flow in a forward-biased PN junction.

The current through a forward-biased diode depends exponentially upon the applied voltage (Figure 1.13). About 0.6V suffices to produce substantial forward conduction in a silicon PN junction at room temperature.[5] Because diffusion is caused by the thermal motions of carriers, higher temperatures cause an exponential increase in diffusion currents. The forward current through a PN junction thus increases exponentially as temperature increases. Expressed another way, the forward bias required to sustain a constant current in a silicon PN junction decreases by approximately 2mV/°C.

Figure 1.13 also shows a low level of current flow when the diode is reverse-biased. This current flow is called *reverse conduction* or *leakage*. Leakage currents are produced by the few minority carriers thermally generated in the silicon. The electric field opposes the flow of majority carriers across a reverse-biased junction,

[5] The most widely quoted value is 0.7V, but in practice a typical integrated circuit base-emitter junction under microamp-level bias at 25°C exhibits a value nearer 0.6V than 0.7V.

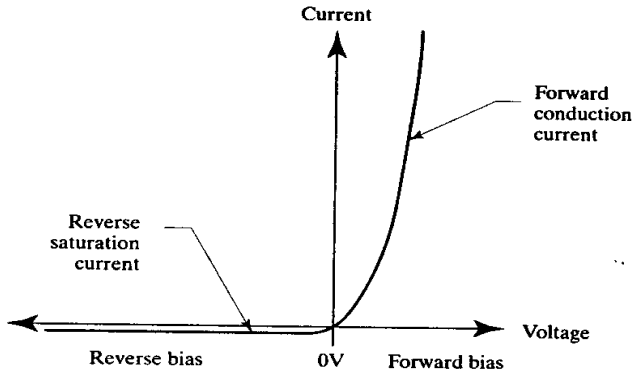


FIGURE 1.13 Diode conduction characteristics. The current scale is greatly magnified to show the reverse saturation current, which typically equals no more than a few picoamps at 25°C.

but it aids the flow of minority carriers. The application of a reverse bias sweeps these minority carriers across the junction. Because the rate of generation of minority carriers in the bulk silicon is essentially independent of electric fields, the leakage current does not vary much with reverse bias. Thermal generation does increase with temperature, and leakage currents are therefore temperature-dependent. In silicon, leakage currents double approximately every eight degrees Celsius. At high temperatures, the leakage currents begin to approach the operating currents of the circuit. The maximum operating temperature of a semiconductor device is therefore limited by leakage current. A maximum junction temperature of 150°C is widely accepted for silicon-integrated circuits.[6]

1.2.3. Schottky Diodes

Rectifying junctions can also form between a semiconductor and a metal. Such junctions are called *Schottky barriers*. The behavior of a Schottky barrier is somewhat analogous to that of a PN junction. Schottky barriers can, for example, be used to construct *Schottky diodes*, which behave much like PN diodes. Schottky barriers can also form in the contact regions of an integrated circuit's interconnection system.

The *work function* of a material equals the amount of energy required to remove an electron from it. Each material has its own characteristic work function that depends upon the properties of its crystal lattice as well as its composition. When two materials with different work functions are brought into contact, the electrons in each material have different initial energies. A voltage difference called the *contact potential* therefore exists between the two materials. Consider the case of a PN junction. The two semiconductors on either side of the junction have the same lattice structure. The contact potential of a PN junction, or its *built-in potential*, depends only upon doping. In the case of a Schottky barrier, the different lattice structures of the metal and the semiconductor also contribute to the contact potential.

A typical rectifying Schottky barrier results when aluminum metal touches lightly doped N-type silicon (Figure 1.14B). The carriers must redistribute in order to counterbalance the contact potential. Electrons diffuse from the semiconductor into the metal, where they pile up to form a thin film of negative charge. This exodus of electrons from the silicon leaves behind a zone of ionized dopant atoms that form a

[6] Integrated circuits can be built that work at 200°C, but many standard design practices do not apply. See R. J. Widlar and M. Yamatake, "Dynamic Safe-Area Protection for Power Transistors Employs Peak-Temperature Limiting." *IEEE J. Solid-State Circuits*, SC-22, #1, 1987, p. 77–84.

depletion region (Figure 1.14A). The electric field generated by the depletion region draws electrons from the metal back into the semiconductor. Equilibrium occurs when the drift and diffusion currents are equal. The potential difference across the Schottky barrier now equals the contact potential. Few minority carriers exist on the semiconductor side of the Schottky barrier, so the Schottky diode is called a *majority-carrier device.*

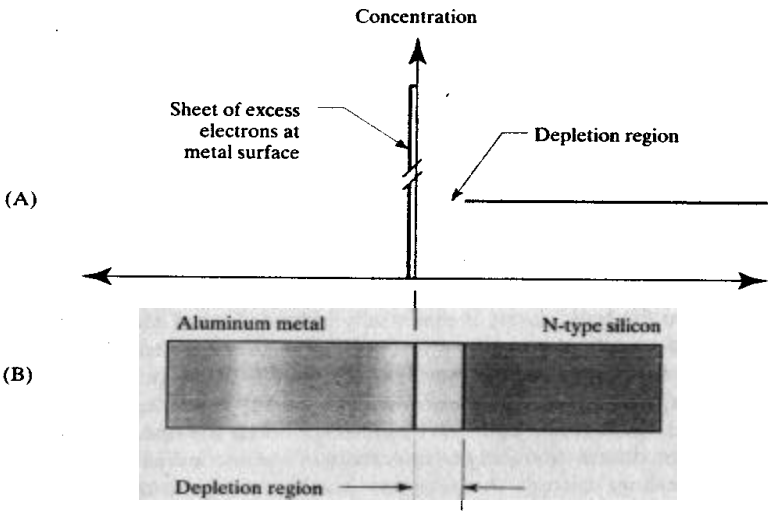


FIGURE 1.14 Diagram of excess carrier concentrations on either side of the Schottky barrier (A) and cross-section of the corresponding Schottky structure (B).

The behavior of a Schottky diode under bias can be similarly analyzed. The N-type silicon forms the *cathode* of the diode, while the metal plate forms the *anode.* The case of a zero-biased Schottky diode is identical to the case of the equilibrium Schottky barrier analyzed above. A reverse-biased Schottky has an external voltage connected in order to bias the semiconductor positively with respect to the metal. The resulting voltage difference adds to the contact potential. The depletion region widens to counterbalance the increased voltage difference, equilibrium is restored, and very little current flows through the diode.

A forward-biased Schottky diode has an external voltage connected in order to bias the metal positively with respect to the semiconductor. The resulting voltage difference across the junction opposes the contact potential, and the width of the depletion region shrinks. Eventually the contact potential is entirely offset, and a depletion region attempts to form on the metal side of the junction. The metal, being a conductor, cannot support an electric field, and no depletion region can form to oppose the externally applied potential. This potential begins to sweep electrons across the junction from the semiconductor into the metal, and a current flows through the diode.

Schottky diodes exhibit current-voltage characteristics similar to those of a PN diode (Figure 1.13). Schottky diodes also exhibit leakage currents caused by low levels of minority carrier injection from the metal into the semiconductor. These conduction mechanisms are accelerated by high temperatures, producing temperature dependencies similar to those of a PN diode.

Despite many apparent similarities, there are a few fundamental differences between Schottky diodes and PN junction diodes. Schottky diodes are majority-carrier devices since they rely primarily upon majority-carrier conduction. At high current densities a few holes do flow from the metal to the semiconductor, but

these contribute only a small fraction of the total current. Schottky diodes do not support large excess minority-carrier populations. Since the switching speed of a diode is a function of the time required for excess minority carriers to recombine, Schottky diodes can switch very rapidly. Some types of Schottky diodes also exhibit lower forward bias voltages than PN diodes. This combination of low forward-voltage drop and highly efficient switching make Schottky diodes very useful devices.

Schottky diodes can also be formed to P-type silicon, but the forward biases required for conduction are usually quite low. This renders P-type Schottky diodes rather leaky and they are therefore rarely used.[7] Most practical Schottky diodes result from the union between lightly doped N-type silicon and a class of materials called *silicides*. These substances are definite compounds of silicon and certain metals, for example platinum and palladium. Silicides exhibit very stable work functions and therefore form Schottky diodes that have consistent and repeatable characteristics.

1.2.4. Zener Diodes

Under normal conditions, only a small current flows through a reverse-biased PN junction. This leakage current remains approximately constant until the reverse bias exceeds a certain critical voltage, beyond which the PN junction suddenly begins to conduct large amounts of current (Figure 1.15). The sudden onset of significant reverse conduction is called *reverse breakdown,* and it can lead to device destruction if the current flow is not limited by some external means. Reverse breakdown often sets the maximum operating voltage of a solid-state device. However, if appropriate precautions are taken to limit the current flow, a junction in reverse breakdown can provide a fairly stable voltage reference.

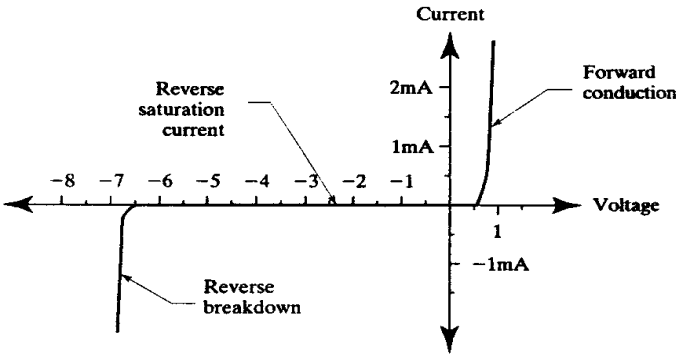


FIGURE 1.15 Reverse breakdown in a PN junction diode.

One of the mechanisms responsible for reverse breakdown is called *avalanche multiplication.* Consider a PN junction under reverse bias. The width of the depletion region increases with bias, but not fast enough to prevent the electric field from intensifying. The intense electric field accelerates the few carriers crossing the depletion region to extremely high velocities. When these carriers collide with lattice atoms, they knock loose valence electrons and generate additional carriers. This

[7] For example, compare the differences in work functions for platinum with respect to N-type silicon (0.85V) and P-type silicon (0.25V): R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons, 1986), p. 157.

process is aptly named because a single carrier can spawn literally thousands of additional carriers through collisions, just as a single snowball can start an avalanche.

The other mechanism behind reverse breakdown is called *tunneling*. Tunneling is a quantum-mechanical process that allows particles to move short distances regardless of any apparent obstacles. If the depletion region is thin enough, then carriers can leap across it by tunneling. The tunneling current depends strongly on both the depletion region width and the voltage difference across the junction. Reverse breakdown caused by tunneling is called *Zener breakdown*.

The reverse breakdown voltage of a junction depends on the width of its depletion region. Wider depletion regions produce higher breakdown voltages. As previously explained, the more lightly doped side of a junction sets its depletion region width and therefore its breakdown voltage. When the breakdown voltage is less than five volts, the depletion region is so thin that Zener breakdown predominates. When the breakdown voltage exceeds five volts, avalanche breakdown predominates. A PN diode designed to operate in reverse conduction is called either a *Zener diode* or an *avalanche diode*, depending on which of these two mechanisms predominates. Zener diodes have breakdown voltages of less than five volts, while avalanche diodes have breakdown voltages of more than five volts. Engineers traditionally call all breakdown diodes Zeners regardless of what mechanism underlies their operation. This can lead to confusion because a 7V Zener conducts primarily by avalanche breakdown.

In practice, the breakdown voltage of a junction depends on its geometry as well as its doping profile. The above discussion analyzed a *planar junction* consisting of two uniformly doped semiconductor regions intersecting in a planar surface. Although some real junctions approximate this ideal, most have curved sidewalls. The curvature intensifies the electric field and reduces the breakdown voltage. The smaller the radius of curvature, the lower the breakdown voltage. This effect can have a dramatic impact on the breakdown voltages of shallow junctions. Most Schottky diodes have sharp discontinuities at the edge of the metal-silicon interface. Electric field intensification can drastically reduce the measured breakdown voltage of a Schottky diode unless special precautions are taken to relieve the electric field at the edges of the Schottky barrier.

Figure 1.16 shows schematic symbols for all of the diodes discussed above. The PN junction diode uses a straight line to denote the cathode, while the Schottky diode and Zener diode are indicated by modifications to the cathode bar. In all cases, the arrow indicates the direction of conventional current flow through the forward-biased diode. In the case of the Zener diode, this arrow can be somewhat misleading because Zeners are normally operated in reverse bias. To the casual observer, the symbol may thus appear to be inserted "the wrong way around."

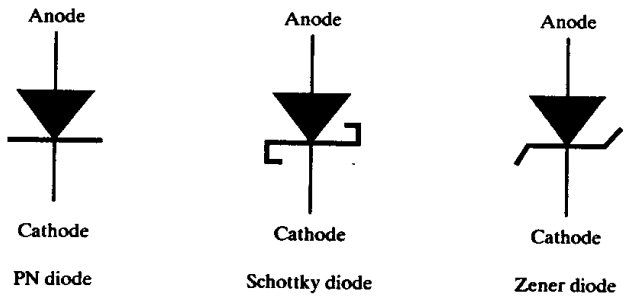


FIGURE 1.16 Schematic symbols for PN junction, Schottky, and Zener diodes. Some schematics show the arrowheads unfilled or show only half the arrowheads.

1.2.5. Ohmic Contacts

Contacts must be made between metals and semiconductors in order to connect solid-state devices into a circuit. These contacts would ideally be perfect conductors, but in practice they are *Ohmic contacts* that exhibit a small amount of resistance. Unlike rectifying contacts, these Ohmic contacts will conduct current equally well in either direction.

Schottky barriers can exhibit Ohmic conduction if the semiconductor material is doped heavily enough. The high concentration of dopant atoms thins the depletion region to the point where carriers can easily tunnel across it. Unlike normal Zener diodes, Ohmic contacts can support tunneling at very low voltages. Rectification does not occur since the carriers can effectively bypass the Schottky barrier by tunneling through it.

An Ohmic contact can also form if a Schottky barrier's contact potential causes surface accumulation rather than surface depletion. In accumulation, a thin layer of majority carriers forms at the semiconductor surface. In the case of an N-type semiconductor, this layer consists of excess electrons. The metal is a conductor and therefore cannot support a depletion region. A thin film of charge thus appears at the surface of the metal to counterbalance the accumulated carriers in the silicon. The lack of a depletion region on either side of the barrier prevents the contact from supporting a voltage differential, and any externally applied voltage will sweep carriers across the junction. Carriers can flow in either direction, so this type of Schottky barrier forms an Ohmic contact rather than a rectifying one.

In practice, rectifying contacts form to lightly doped silicon, and Ohmic contacts form to heavily doped silicon. The exact mechanism behind Ohmic conduction is unimportant since all Ohmic contacts behave in essentially the same manner. A lightly doped silicon region can be Ohmically contacted only if a thin layer of more heavily doped silicon is placed beneath the contact. Contact resistances of less than $50\Omega/\mu m^2$ can be obtained if a heavily doped silicon layer is used in combination with a suitable metal system. This resistance is small enough that it can be neglected for most applications.

Any junction between dissimilar materials exhibits a contact potential equal to the difference between the work functions of the materials. This rule applies to Ohmic contacts as well as to PN junctions and rectifying Schottky barriers. If all the contacts and junctions are held at the same temperature, then the sum of the contact potentials around any closed loop will equal zero. Contact potentials are, however, strong functions of temperature. If one of the junctions is held at a different temperature than the others, then its contact potential will shift and the sum of the contact potentials will no longer equal zero. This *thermoelectric effect* has significant implications for integrated circuit design.

Figure 1.17 shows a block of N-type silicon contacted on either side by aluminum. If one end of the block is heated, then a measurable voltage develops across the

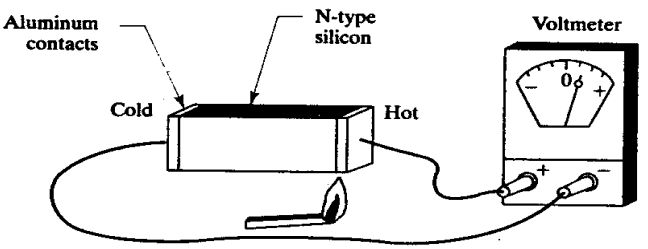


FIGURE 1.17 The thermoelectric effect produces a net measurable voltage if the two contacts are held at different temperatures.

block due to the mismatch between the two contact potentials. This voltage drop is typically 0.1–1.0mV/°C.[8] Many integrated circuits rely upon voltages matching within a millivolt or two, so even small temperature differences are enough to cause such circuits to malfunction.

1.3 BIPOLAR JUNCTION TRANSISTORS

While diodes are useful devices, they cannot amplify signals, and almost all electronic circuits require amplification in one form or another. One device that can amplify signals is called a *bipolar junction transistor* (BJT).

The structures of the two types of bipolar junction transistors are shown in Figure 1.18. Each transistor consists of three semiconductor regions called the *emitter, base,* and *collector.* The base is always sandwiched between the emitter and the collector. An NPN transistor consists of an N-type emitter, a P-type base, and an N-type collector. Similarly, a PNP transistor consists of a P-type emitter, an N-type base, and a P-type collector. In these simplified cross-sections, each region of the transistor consists of a uniformly doped section of a rectangular bar of silicon. Modern bipolar transistors have somewhat different cross-sections, but the principles of operation remain the same.

Figure 1.18 also shows the symbols for the two types of transistors. The arrowhead placed on the emitter lead indicates the direction of conventional current flow through the forward-biased emitter-base junction. No arrow appears on the collector lead even though a junction also exists between the collector and the base. In the simplified transistors of Figure 1.18, the emitter-base and collector-base junctions appear to be identical. One could apparently swap the collector and emitter leads without affecting the behavior of the device. In practice, the two junctions have different doping profiles and geometries and are not interchangeable. The emitter lead is distinguished from the collector lead by the presence of the arrowhead.

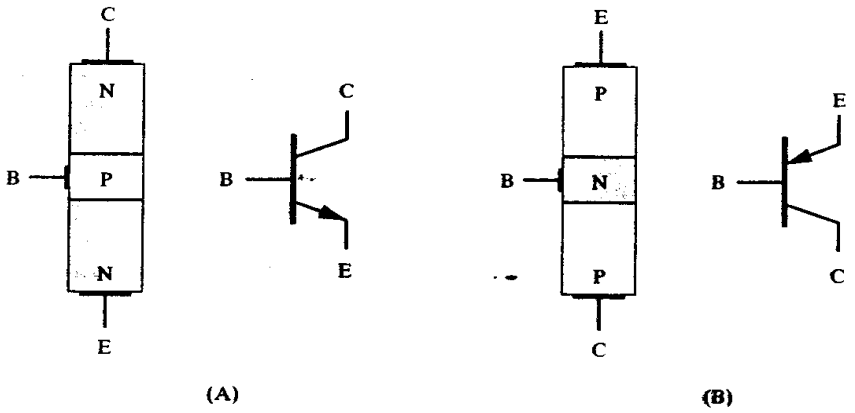


FIGURE 1.18 Structures and schematic symbols for the NPN transistor (A) and the PNP transistor (B).

A bipolar junction transistor can be viewed as two PN junctions connected back-to-back. The base region of the transistor is very thin (about 1–2μm wide). When the two junctions are placed in such close proximity, carriers can diffuse from one junction to the other before they recombine. Conduction across one junction therefore affects the behavior of the other junction.

[8] Lightly-doped silicon exhibits a higher Seebeck voltage; these values are taken from Widlar, *et al.*, p. 79.

Figure 1.19A shows an NPN transistor with zero volts applied across the base-emitter junction and five volts applied across the base-collector junction. Neither junction is forward biased, so very little current flows through any of the three terminals of the transistor. A transistor with both junctions reverse biased is said to be in *cutoff*. Figure 1.19B shows the same transistor with ten microamps of current injected into its base. This current forward biases the base-emitter junction to a potential of about 0.65V. A collector current a hundred times larger than the base current flows across the base-collector junction even though this junction remains reverse biased. This current is a consequence of the interaction between the forward-biased base-emitter junction and the reverse-biased base-collector junction. Whenever a transistor is biased in this manner, it is said to operate in the *forward active* region. If the emitter and collector terminals are interchanged so that the base-emitter junction becomes reverse-biased and the base-collector junction becomes forward-biased, the transistor is said to operate in the *reverse active* region. In practice, transistors are seldom operated in this manner.

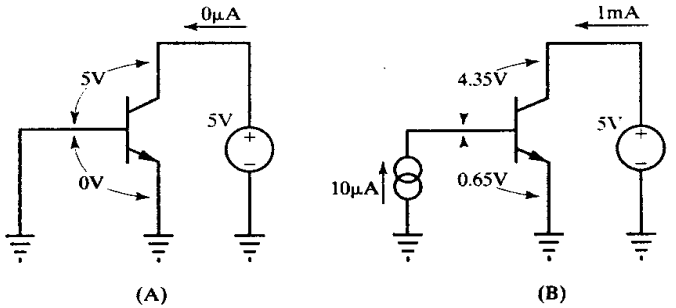


FIGURE 1.19 An NPN transistor operating in cutoff (A) and in the forward active region (B).

Figure 1.20 helps explain why collector current flows across a reverse-biased junction. Carriers flow across the base-emitter junction as soon as it becomes forward biased. Most of the current flowing across this junction consists of electrons injected from the heavily doped emitter into the lightly doped base. Most of these electrons diffuse across the narrow base before they recombine. The base-collector junction is reverse biased, so very few majority carriers can flow from the base into the collector. The same electric field that opposes the flow of majority carriers actually aids the flow of minority carriers. The electrons are minority carriers in the base, so they are swept across the reverse-biased base-collector junction into the collector. Here they again become majority carriers flowing toward the collector terminal. The collector current consists of the electrons that successfully complete the journey from emitter to collector without recombining in the base.

Some of the electrons injected into the base do not reach the collector. Those that do not reach the collector recombine in the base region. Base recombination consumes holes that are replenished by a current flowing in from the base terminal. Some holes are also injected from the base into the emitter, where they rapidly recombine. These holes represent a second source of base terminal current. These recombination processes typically consume no more than 1% of the emitter current, so only a small base current is required to maintain the forward bias across the base-emitter junction.

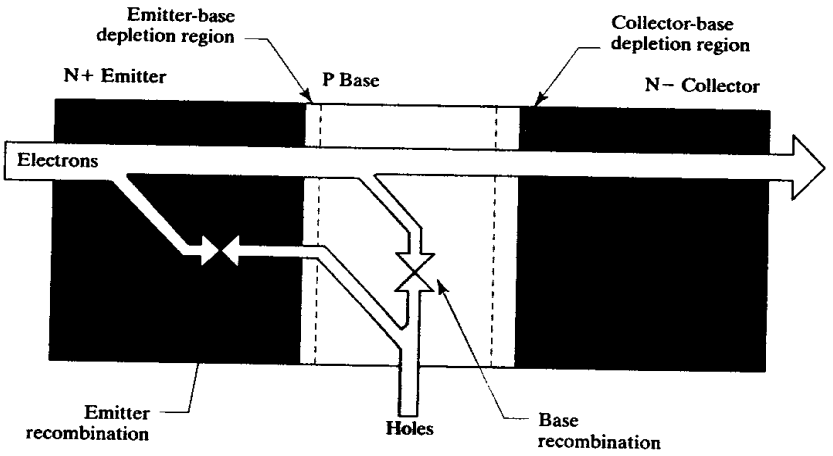


FIGURE 1.20 Current flow in an NPN transistor in the forward-active region.

1.3.1. Beta

The current amplification achieved by a transistor equals the ratio of its collector current to its base current. This ratio has been given various names, including *current gain* and *beta.* Likewise, different authors have used different symbols for it, including β and h_{FE}. A typical integrated NPN transistor exhibits a beta of about 150. Certain specialized devices may have betas exceeding 10,000. The beta of a transistor depends upon the two recombination processes illustrated in Figure 1.20.

Base recombination occurs primarily within the portion of the base between the two depletion regions, which is called the *neutral base* region. Three factors influence the base recombination rate: neutral base width, base doping, and the concentration of recombination centers. A thinner neutral base reduces the distance that the minority carriers must traverse and thus lessens the probability of recombination. Similarly, a more lightly doped base region minimizes the probability of recombination by reducing the majority carrier concentration. The *Gummel number* Q_B measures both of these effects. It is calculated by integrating the dopant concentration along a line traversing the neutral base region. In the case of uniform doping, the Gummel number equals the product of the base dopant concentration and the width of the neutral base. Beta is inversely proportional to the Gummel number.

The switching speed of transistors depends primarily on how quickly the excess minority carriers can be removed from the base, either through the base terminal or through recombination. Gold-doping is sometimes used to deliberately increase the number of recombination centers in bipolar junction transistors. The elevated recombination rate helps speed transistor switching, but it also reduces transistor beta. Few analog integrated circuits are built on gold-doped processes because of their low betas.

Bipolar transistors typically use a lightly doped base and a heavily doped emitter. This combination helps ensure that almost all of the current injected across the base-emitter junction consists of carriers flowing from emitter to base and not *vice-versa.* Heavy doping enhances the recombination rate in the emitter, but this has little impact since so few carriers are injected into the emitter in the first place. The

ratio of current injected into the emitter to that injected into the base is called the *emitter injection efficiency.*

Most NPN transistors use a wide, lightly doped collector in combination with a heavily doped emitter and a thin. moderately doped base. The light collector doping allows a wide depletion region to form in the neutral collector. This permits a high collector operating voltage without avalanching the collector-base junction. The asymmetric doping of emitter and collector helps explain why bipolar transistors do not operate well when these terminals are swapped.[9] A typical integrated NPN transistor with a forward beta of 150 has a reverse beta of less than 5. This difference is primarily due to the drastic reduction in emitter injection efficiency caused by the substitution of a lightly doped collector for a heavily doped emitter.

Beta also depends upon collector current. Beta is reduced at low currents by leakage and by low levels of recombination in the depletion regions. At modest current levels, these effects become insignificant and the beta of the transistor climbs to a peak value determined by the mechanisms discussed above. High collector currents cause beta to roll off due to an effect called *high-level injection.* When the minority carrier concentration approaches the majority carrier concentration in the base, extra majority carriers accumulate to maintain the balance of charges. The additional base majority carriers cause the emitter injection efficiency to decrease, which in turn causes beta to decrease. Most transistors are operated at moderate current levels to avoid beta roll-off, but power transistors must often operate in high-level injection because of size constraints.

The behavior of the PNP transistor is very similar to that of the NPN transistor. The beta of a PNP transistor is lower than that of an NPN of comparable dimensions and doping profiles, because the mobility of holes is lower than that of electrons. In many cases, the performance of the PNP is further degraded because of a conscious choice to optimize the NPN transistor at the expense of the PNP. For example, the material used to construct the base region of an NPN is often used to fabricate the emitter of a PNP. Since the resulting emitter is rather lightly doped, emitter injection efficiency is low, and the onset of high-level injection occurs at moderate current levels. Despite their limitations, PNP transistors are very useful devices, and most bipolar processes support their construction.

1.3.2. I-V Characteristics

The performance of a bipolar transistor can be graphically depicted by drawing a family of curves that relate base current, collector current, and collector-emitter voltage. Figure 1.21 shows a typical set of curves for an integrated NPN transistor. The vertical axis measures collector current I_C, while the horizontal axis measures collector-to-emitter voltage V_{CE}. A number of curves are superimposed upon the same graph, each representing a different base current I_B. This family of curves shows a number of interesting features of the bipolar junction transistor.

In the *saturation region,* the collector-emitter voltage remains so small that the collector-base junction is slightly forward-biased. The electric field that sweeps minority carriers across the collector-base junction still exists, so the transistor continues to conduct current. The collector-emitter voltage remains so low that Ohmic resistances in the transistor (particularly those in the lightly doped collector) become significant. The current supported in saturation is therefore less than that supported in the forward active region. The saturation region is of particular interest to integrated circuit

[9] This is only part of the explanation. The effective base width of the transistor also increases when it is operated in the reverse active mode.

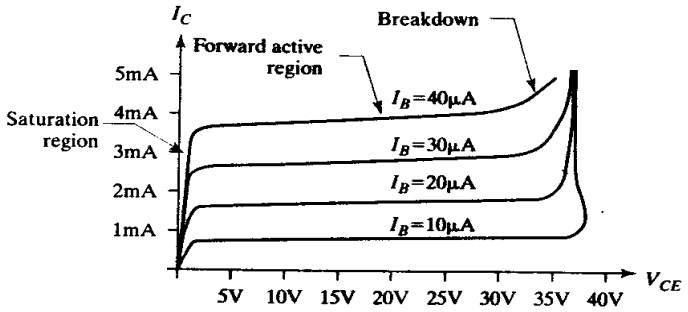


FIGURE 1.21 Typical I-V plot of an NPN transistor.

designers because the forward biasing of the collector-base junction injects minority carriers into the neutral collector. Section 8.1.4 discusses the effects of saturation upon integrated bipolar transistors in greater detail.

The collector-emitter voltage in the forward active region is large enough to reverse bias the collector-base junction. Ohmic drops in the collector no longer significantly reduce the electric field across the collector-base junction, so the current flow through the transistor now depends solely upon beta. The slight upward tilt to the current curves results from the *Early effect.* As the reverse bias on the collector-base junction increases, the depletion region at this junction widens and consequently the neutral base narrows. Since beta depends on base width, it increases slightly as the collector-emitter voltage rises. The Early effect can be minimized by using a very lightly doped collector, so the depletion region extends primarily into the collector rather than into the base.

Beyond a certain collector-emitter voltage, the collector current increases rapidly. This effect limits the maximum operating voltage of the transistor. In the case of a typical integrated NPN transistor, this voltage equals some 30V–40V. The increased current flow results from either one of two effects, the first of which is avalanche breakdown. The collector-base junction will avalanche if it is sufficiently reverse-biased. A wide lightly doped collector region can greatly increase the avalanche voltage rating, and discrete power transistors can achieve operating voltages of more than a thousand volts.

The second limiting mechanism is *base punchthrough.* Punchthrough occurs when the collector-base depletion region reaches all the way through the base and merges with the base-emitter depletion region. Once this occurs, carriers can flow directly from emitter to collector, and current is limited only by the resistance of the neutral collector and emitter. The resulting rapid increase in collector current mimics the effects of avalanche breakdown.

Base punchthrough is often observed in high-gain transistors. For example, *super-beta* transistors use an extremely thin base region to obtain betas of a thousand or more. Base punchthrough limits the operating voltage of these devices to a couple of volts. Super-beta transistors also display a pronounced Early effect because of the encroachment of the collector-base depletion region into the extremely thin neutral base. General-purpose transistors use wider base regions to reduce the Early effect, and their operating voltages are usually limited by avalanche instead of base punchthrough (Section 8.1.2).

1.4 MOS TRANSISTORS

The bipolar junction transistor amplifies a small change in input current to provide a large change in output current. The gain of a bipolar transistor is thus defined as the ratio of output to input current (beta). Another type of transistor, called a *field-*

effect transistor (FET), transforms a change in input voltage into a change in output current. The gain of an FET is measured by its *transconductance*, defined as the ratio of change in output current to change in input voltage.

The field-effect transistor is so named because its input terminal (called its *gate*) influences the flow of current through the transistor by projecting an electric field across an insulating layer. Virtually no current flows through this insulator, so the gate current of a FET transistor is vanishingly small. The most common type of FET uses a thin silicon dioxide layer as an insulator beneath the gate electrode. This type of transistor is called a ***metal-oxide-semiconductor*** (MOS) transistor, or alternatively, a ***metal-oxide-semiconductor*** *field-effect transistor* (MOSFET). MOS transistors have replaced bipolars in many applications because they are smaller and can often operate using less power.

The MOS transistor can be better understood by first considering a simpler device called a *MOS capacitor*. This device consists of two electrodes, one of metal and one of extrinsic silicon, separated by a thin layer of silicon dioxide (Figure 1.22A). The metal electrode forms the *gate*, while the semiconductor slab forms the *backgate* or *body*. The insulating oxide layer between the two is called the *gate dielectric*. The illustrated device has a backgate consisting of lightly doped P-type silicon. The electrical behavior of this MOS capacitor can be demonstrated by grounding the backgate and biasing the gate to various voltages. The MOS capacitor of Figure 1.22A has a gate potential of 0V. The difference in work functions between the metal gate and the semiconductor backgate causes a small electric field to appear across the dielectric. In the illustrated device, this field biases the metal plate slightly positive with respect to the P-type silicon. This electric field attracts electrons from deep within the silicon up toward the surface, while it repels holes away from the surface. The field is weak, so the change in carrier concentrations is small and the overall effect upon the device characteristics is minimal.

Figure 1.22B shows what occurs when the gate of the MOS capacitor is biased positively with respect to the backgate. The electric field across the gate dielectric strengthens and more electrons are drawn up from the bulk. Simultaneously, holes are repelled away from the surface. As the gate voltage rises, a point is reached where more electrons than holes are present at the surface. Due to the excess electrons, the surface layers of the silicon behave as if they were N-type. The apparent reversal of doping polarity is called *inversion* and the layer of silicon that inverts is called a *channel*. As the gate voltage increases still further, more electrons accumulate at the surface and the channel becomes more strongly inverted. The voltage at which the channel just begins to form is called the *threshold voltage* V_t. When the voltage difference between gate and backgate is less than the threshold voltage, no channel forms. When the voltage difference exceeds the threshold voltage, a channel appears.

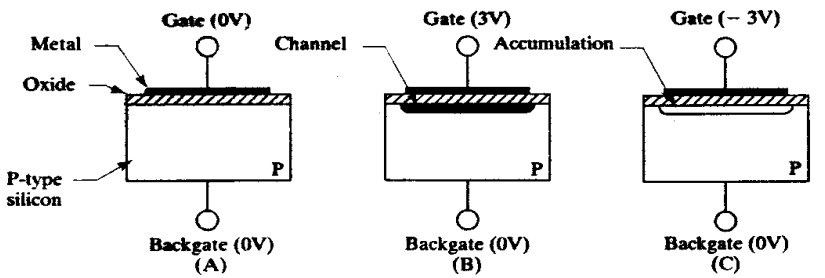


FIGURE 1.22 MOS capacitor: (A) unbiased (V_{BG} = 0V), (B) inversion (V_{BG} = 3V), (C) accumulation (V_{BG} = –3V).

Figure 1.22C shows what happens if the gate of the MOS capacitor is biased negatively with respect to the backgate. The electric field now reverses, drawing holes toward the surface and repelling electrons away from it. The surface layers of silicon appear to be more heavily doped, and the device is said to be in *accumulation.*

The behavior of the MOS capacitor can be utilized to form a true MOS transistor. Figure 1.23A shows the cross section of the resulting device. The gate, dielectric, and backgate remain as before. Two additional regions are formed by selectively doping the silicon on either side of the gate. One of these regions is called the *source* and the other is called the *drain.* Imagine that the source and backgate are both grounded and that a positive voltage is applied to the drain. As long as the gate-to-backgate voltage remains less than the threshold voltage, no channel forms. The PN junction formed between drain and backgate is reverse-biased, so very little current flows from drain to backgate. If the gate voltage exceeds the threshold voltage, a channel forms beneath the gate dielectric. This channel acts like a thin film of N-type silicon shorting the source to the drain. A current consisting of electrons flows from the source across the channel to the drain. In summary, drain current will only flow if the gate-to-source voltage V_{GS} exceeds the threshold voltage V_t.

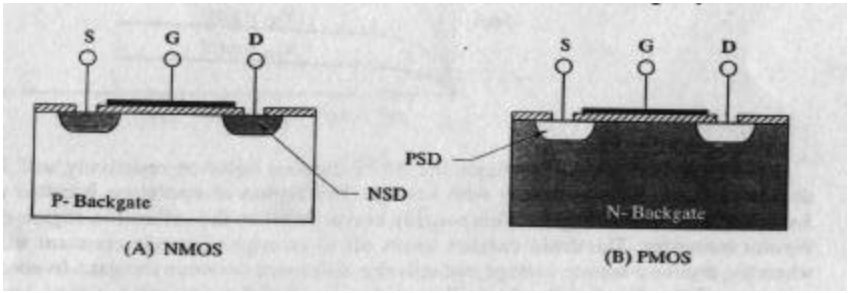


FIGURE 1.23 Cross sections of MOSFET transistors: NMOS (A) and PMOS (B). In these diagrams, S = Source, G = Gate, and D = Drain. The backgate connections, though present, are not illustrated.

The source and drain of a MOS transistor are interchangeable, as both are simply N-type regions formed in the P-type backgate. In many cases, these two regions are identical and the terminals can be reversed without changing the behavior of the device. Such a device is said to be *symmetric.* In a symmetric MOS transistor the labeling of source and drain becomes somewhat arbitrary. By definition, carriers flow out of the source and into the drain. The identity of the source and the drain therefore depends on the biasing of the device. Sometimes the bias applied across the transistor fluctuates and the two terminals swap roles. In such cases, the circuit designer must arbitrarily designate one terminal the drain and the other the source.

Asymmetric MOS transistors are designed with different source and drain dopings and geometries. There are several reasons why transistors may be made asymmetric, but the result is the same in every case. One terminal is optimized to function as the drain and the other as the source. If source and drain are swapped, then the performance of the device will suffer.

The transistor depicted in Figure 1.23A has an N-type channel and is therefore called an *N-channel MOS transistor,* or NMOS. *P-channel MOS* (PMOS) transistors also exist. Figure 1.23B shows a sample PMOS transistor consisting of a lightly doped N-type backgate with P-type source and drain regions. If the gate of this transistor is biased positive with respect to the backgate, then electrons are drawn to the surface and holes are repelled away from it. The surface of the silicon accumulates, and no channel forms. If the gate is biased negative with respect to the backgate, then holes are drawn to the surface, and a channel forms. The PMOS transistor thus

has a negative threshold voltage. Engineers often ignore the sign of the threshold voltage since it is normally positive for NMOS transistors and negative for PMOS transistors. An engineer might say, "The PMOS V_t has increased from 0.6V to 0.7V" when in actuality the PMOS V_t has shifted from –0.6V to –0.7V.

1.4.1. Threshold Voltage

The *threshold voltage* of a MOS transistor equals the gate-to-source bias required to just form a channel with the backgate of the transistor connected to the source. If the gate-to-source bias is less than the threshold voltage, then no channel forms. The threshold voltage exhibited by a given transistor depends on a number of factors, including backgate doping. dielectric thickness, gate material, and excess charge in the dielectric. Each of these effects will be briefly examined.

Backgate doping has a major effect on the threshold voltage. If the backgate is doped more heavily, then it becomes more difficult to invert. A stronger electric field is required to achieve inversion, and the threshold voltage increases. The backgate doping of an MOS transistor can be adjusted by performing a shallow implant beneath the surface of the gate dielectric to dope the channel region. This type of implant is called a *threshold adjust implant* (or *V_t adjust implant*).

Consider the effects of a V_t adjust implant upon an NMOS transistor. If the implant consists of acceptors, then the silicon surface becomes more difficult to invert and the threshold voltage increases. If the implant consists of donors, then the surface becomes easier to invert and the threshold decreases. If enough donors are implanted, the surface of the silicon can actually become counterdoped. In this case, a thin layer of N-type silicon forms a permanent channel at zero gate bias. The channel becomes more strongly inverted as the gate bias increases. As the gate bias is decreased, the channel becomes less strongly inverted and at some point it vanishes. The threshold voltage of this NMOS transistor is actually negative. Such a transistor is called a *depletion-mode NMOS*, or simply a *depletion NMOS*. In contrast, an NMOS with a positive threshold voltage is called an *enhancement-mode NMOS*, or *enhancement NMOS*. The vast majority of commercially fabricated MOS transistors are enhancement-mode devices, but there are a few applications that require depletion-mode devices. A depletion-mode PMOS can also be constructed. Such a device will have a positive threshold voltage.

Depletion-mode devices should always be explicitly identified as such. One cannot rely on the sign of the threshold voltage to convey this information, because many engineers customarily ignore threshold polarities. Therefore, one should say "a depletion-mode PMOS with a threshold of 0.7V," rather than a PMOS with a threshold of 0.7V." Many engineers would interpret the latter statement as indicating an enhancement PMOS with a threshold of –0.7V rather than a depletion PMOS with a threshold of +0.7V. Explicitly referring to depletion-mode devices as such eliminates any possibility of confusion.

Special symbols are often used to distinguish between different types of MOS transistors. Figure 1.24 shows a representative collection of these symbols.[10] Symbols A and B are the standard symbols for NMOS and PMOS transistors, respectively. These symbols are not commonly used in the industry; instead symbols

[10] Symbols A, B, E, F, G, and H are used by various authors; see A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design* (New York: John Wiley and Sons, 1984), pp. 112–113; also P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed. (New York: John Wiley and Sons, 1993), p. 60. The *J. Solid State Circuits* also uses three-terminal MOS symbols but differentiates PMOS devices by placing a bubble on their gate leads.

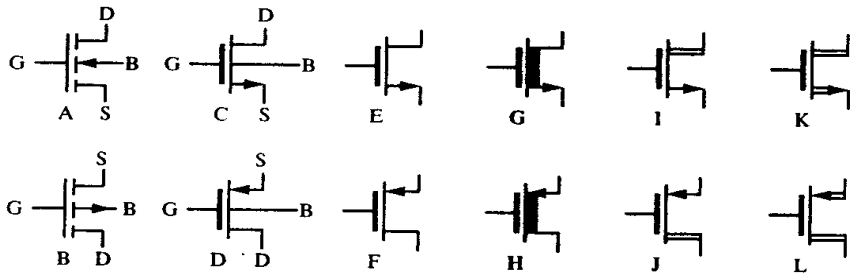


FIGURE 1.24 MOSFET symbols: A, B: standard symbols; C, D: industry symbols (four-terminal); E, F: industry symbols (three-terminal); G, H: depletion-mode devices; I, J: asymmetric high-voltage MOS symbols; K, L: symmetric high-voltage MOS symbols.

C and D are preferred for NMOS and PMOS transistors, respectively. These symbols intentionally resemble NPN and PNP transistors. This convention helps highlight the essential similarities between MOS and bipolar circuits. Symbols E and F are sometimes employed when the backgates of the transistors connect to known potentials. Every MOS transistor has a backgate, so this terminal must always connect to something. Symbols E and F are potentially confusing, because the reader must infer the backgate connections. These symbols are nonetheless very popular because they make schematics much more legible. Symbols G and H are often used for depletion-mode devices, where the solid bar from drain to source represents the channel present at zero bias. Symbols I and J are sometimes employed for asymmetric transistors with high-voltage drains, and symbols K and L are used for symmetric transistors with high-voltage terminations for both source and drain. There are many other schematic symbols for MOS transistors; the ones shown in Figure 1.24 form only a representative sample.

Returning to the discussion of threshold voltage, the dielectric also plays an important role in determining the threshold voltage. A thicker dielectric weakens the electric field by separating the charges by a greater distance. Thus, thicker dielectrics increase the threshold voltage while thinner ones reduce it. In theory, the material of the dielectric also affects the electric field strength. In practice, almost all MOS transistors use pure silicon dioxide as the gate dielectric. This material can be grown in extremely thin films of exceptional purity and uniformity; no other material has comparable properties. Alternate dielectric materials therefore have very limited application.[11]

The gate electrode material also affects the threshold voltage of the transistor. As mentioned above, an electric field appears across the gate oxide when the gate and backgate are shorted together. This field is produced by the difference in work functions between the gate and backgate materials. Most practical transistors use heavily doped polysilicon for the gate electrode. The work function of polysilicon can be varied to a limited degree by changing its doping.

A potentially troublesome source of threshold voltage variation comes from the presence of excess charges in the gate oxide or along the interfaces between the oxide and the silicon surface. These charges may consist of ionized impurity atoms, trapped carriers, or structural defects. The presence of trapped electric charge in the dielectric or along its interfaces alters the electric field and therefore the threshold voltage. If the amount of trapped charge varies with time, temperature, or applied bias, then the threshold voltage will also vary. This subject is discussed in greater detail in Section 4.2.2.

[11] A few devices have been fabricated using high-permittivity materials such as silicon nitride for the gate dielectric. Some authors use the term *insulated-gate field effect transistor* (IGFET) to refer to all MOS-like transistors, including those with non-oxide dielectrics.

1.4.2. I-V Characteristics

The performance of an MOS transistor can be graphically illustrated by drawing a family of I-V curves similar to those used for bipolar transistors. Figure 1.25 shows a typical set of curves for an enhancement NMOS. The source and backgate were connected together to obtain these particular curves. The vertical axis measures drain current I_D, while the horizontal axis measures drain-to-source voltage V_{DS}. Each curve represents a specific gate-to-source voltage V_{GS}. The general character of the curves resembles that of the bipolar transistor shown in Figure 1.21, but the family of curves for an MOS transistor are obtained by stepping gate voltage, while those for a bipolar transistor are obtained by stepping base current.

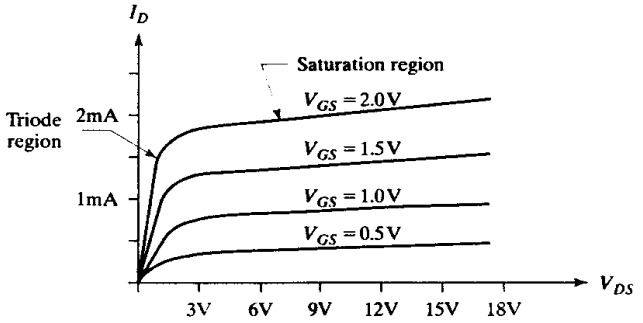


FIGURE 1.25 Typical I-V plot of an NMOS transistor.

At low drain-to-source voltages the MOS channel behaves resistively, and the drain current increases linearly with voltage. This region of operation is called the *linear region* or *triode region*. This roughly corresponds to the saturation region of a bipolar transistor. The drain current levels off to an approximately constant value when the drain-to-source voltage exceeds the difference between the gate-to-source voltage and the threshold voltage. This region is called the *saturation region*, and it roughly corresponds to the forward active region of a bipolar transistor. The term *saturation* thus has very different meanings for MOS and bipolar transistors.

The behavior of the MOS transistor in the linear region is easily explained. The channel acts as a film of doped silicon with a characteristic resistance that depends upon the carrier concentration. The current increases linearly with voltage, exactly as one would expect of a resistor. Higher gate voltages produce larger carrier concentrations and therefore lessen the resistance of the channel. PMOS transistors behave similarly to NMOS transistors, but since holes have lower mobilities than electrons, the apparent resistance of the channel is considerably greater. The effective resistance of an MOS transistor operating in the triode region is symbolized $R_{DS(on)}$.

MOS transistors saturate because of a phenomenon called *pinch-off*. While the drain-to-source voltage remains small, a depletion region of uniform thickness surrounds the channel (Figure 1.26A). As the drain becomes more positive with respect to the source, the depletion region begins to thicken at the drain end. This depletion region intrudes into the channel and narrows it. Eventually the channel depletes all the way through and it is said to have *pinched off* (Figure 1.26B). Carriers move down the channel propelled by the relatively weak electric field along it. When they reach the edge of the pinched-off region, they are sucked across the depletion region by the strong electric field. The voltage drop across the channel does not increase as the drain voltage is increased; instead the pinched-off region widens. Thus, the drain current reaches a limit and ceases to increase.

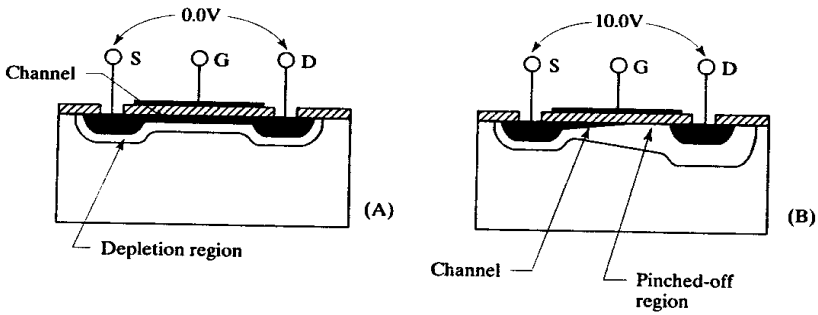


FIGURE 1.26 Behavior of a MOS transistor under bias: (A) V_{DS} = 0V (triode region); (B) V_{DS} = 10V (saturation region).

The drain current curves actually tilt slightly upward in the saturation region. This tilt is caused by *channel length modulation*, which is the MOS equivalent of the Early effect. Increases in drain voltage cause the pinched-off region to widen and the channel length to shorten. The shorter channel still has the same potential drop across it, so the electric field intensifies and the carriers move more rapidly. The drain current thus increases slightly with increasing drain-to-source voltage.

The I-V curves of Figure 1.25 were obtained with the backgate of the transistor connected to the source. If the backgate is biased independently of the source, then the apparent threshold voltage of the transistor will vary. If the source of an NMOS transistor is biased above its backgate, then its apparent threshold voltage increases. If the source of a PMOS transistor is biased below its backgate, then its threshold voltage decreases (it becomes more negative). This *backgate effect*, or *body effect*, arises because the backgate-to-source voltage modulates the depletion region beneath the channel. This depletion region widens as the backgate-to-source differential increases, and it intrudes into the channel as well as into the backgate. A high backgate-to-source differential will thin the channel, which in turn raises the apparent threshold voltage. The intrusion of the depletion region into the channel becomes more significant as the backgate doping rises, and this in turn increases the magnitude of the body effect.

MOS transistors are normally considered majority carrier devices, which conduct only after a channel forms. This simplistic view does not explain the low levels of conduction that occur at gate-to-source voltages just less than the threshold voltage. The formation of a channel is a gradual process. As the gate-to-source voltage increases, the gate first attracts small numbers of minority carriers to the surface. The concentration of minority carriers rises as the voltage increases. When the gate-to-source voltage exceeds the threshold, the number of minority carriers becomes so large that the surface of the silicon inverts and a channel forms. Before this occurs, minority carriers can still move from the source to the drain by diffusion. This *subthreshold conduction* produces currents that are much smaller than those that would flow if a channel were present. However, they are still many orders of magnitude greater than junction leakages. Subthreshold conduction is typically significant only when the gate-to-source voltage is within about 0.3V of the threshold voltage. This is sufficient to cause serious "leakage" problems in low-V_t devices. Some electrical circuits actually take advantage of the exponential voltage-to-current relationship of subthreshold conduction, but these circuits cannot operate at temperatures much in excess of 100°C because the junction leakages become so large that they overwhelm the tiny subthreshold currents.

As with bipolar transistors, MOS transistors can break down by either avalanche or punchthrough. If the voltage across the depletion region at the drain becomes so large that avalanche multiplication occurs, the drain current increases rapidly. Similarly, if the entire channel pinches off, then the source and drain will be shorted by the resulting depletion region and the transistor will punch through.

The operating voltage of an MOS transistor is often limited to a value considerably below the onset of avalanche or punchthrough by a long-term degradation mechanism called *hot carrier injection.* Carriers that traverse the pinched-off portion of the drain are accelerated by the strong electric field present here. The carriers can achieve velocities far beyond those normally associated with room-temperature thermal diffusion, so they are called *hot carriers.* When these carriers collide with atoms near the silicon surface, some of them are deflected up into the gate oxide, and a few of these become trapped. Slowly, over a long period of operation, the concentration of these trapped carriers increases and the threshold voltage shifts. Hot hole injection occurs less readily than hot electron injection because the lower mobility of holes limits their velocity and therefore their ability to surmount the oxide interface. For this reason, NMOS transistors are frequently limited to lower operating voltages than PMOS transistors of similar construction. Various techniques have been devised to limit hot carrier injection (Section 12.1).

1.5 JFET TRANSISTORS

The MOS transistor represents only one type of field-effect transistor. Another is the *junction field-effect transistor* or JFET. This device uses the depletion regions surrounding reverse-biased junctions as a gate dielectric. Figure 1.27A shows a cross-section of an N-channel JFET. This device consists of a bar of lightly doped N-type silicon called the *body* into which two P-type diffusions have been driven from opposite sides. The thin region of N-type silicon remaining between the junctions forms the *channel* of the JFET. The two diffusions act as the *gate* and the *backgate* and the opposite ends of the body form the *source* and the *drain.*

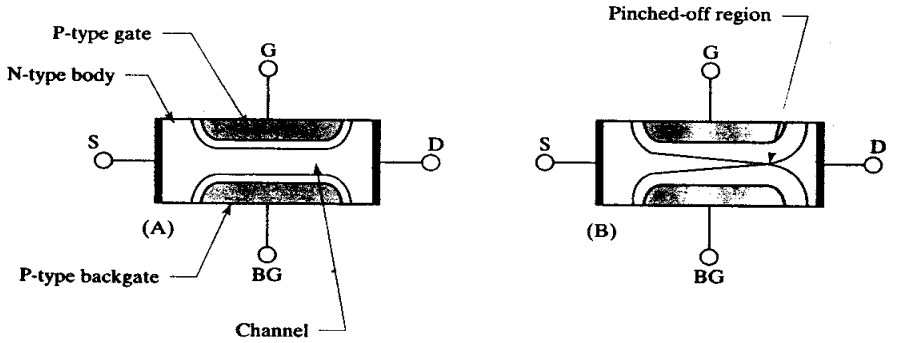


FIGURE 1.27 Cross sections of an N-channel JFET transistor operating in the linear region (A) and in saturation (B). In both diagrams, S = Source, D = Drain, G = Gate, and BG = Backgate.

Suppose that all four terminals of the N-channel JFET are grounded. Depletion regions form around the gate-body and backgate-body junctions. These depletion regions extend into the lightly doped channel, but they do not actually touch one another. A channel therefore exists from the drain to the source. If the drain voltage rises above the source voltage, then a current flows through the channel from

drain to source. The magnitude of this current depends on the resistance of the channel, which in turn depends on its dimensions and doping. As long as the drain-to-source voltage remains small, it does not significantly alter the depletion regions bounding the channel. The resistance of the channel therefore remains constant and the drain-to-source voltage varies linearly with drain current. Under these conditions, the JFET is said to operate in its *linear region*. This region of operation corresponds to the linear (or triode) region of an MOS transistor. Since a channel forms at $V_{GS} = 0$, the JFET resembles a depletion-mode MOSFET rather than an enhancement-mode one.

The depletion regions at the drain end of the JFET widen as the drain voltage increases. The channel becomes increasingly constricted by the encroachment of the opposing depletion regions. Eventually the depletion regions meet and pinch off the channel (Figure 1.27B). Drain current still flows through the transistor even though the channel has pinched off. This current originates at the source terminal and consists of majority carriers (electrons). These carriers move down the channel until they reach the pinched-off region. The large lateral electric field across this region draws the carriers across into the neutral drain.

Further increases in drain voltage have little effect once the channel has pinched off. The pinched-off region widens slightly, but the dimensions of the channel remain about the same. The resistance of the channel determines the magnitude of the drain current, so this also remains approximately constant. Under these conditions, the JFET is said to operate in *saturation*.

The gate and backgate electrodes also influence the current that flows through the channel. As magnitudes of the gate-body and backgate-body voltages increase, the reverse biases across the gate-body and backgate-body junctions slowly increase. The depletion regions that surround these junctions widen and the channel constricts. Less current can flow through the constricted channel, and the drain-to-source voltage required to pinch the channel off decreases. As the magnitudes of the gate and backgate voltages continue to increase, eventually the channel will pinch off even at $V_{DS} = 0$. Once this occurs, no current can flow through the transistor regardless of drain-to-source voltage, and the transistor is said to operate in *cutoff*.

Figure 1.28 shows the I-V characteristics of an N-channel JFET whose gate and backgate electrodes have been connected to one another. Each curve represents a different value of the gate-to-source voltage V_{GS}. The drain currents are at their greatest when $V_{GS} = 0$, and they decrease as the magnitude of the gate voltage

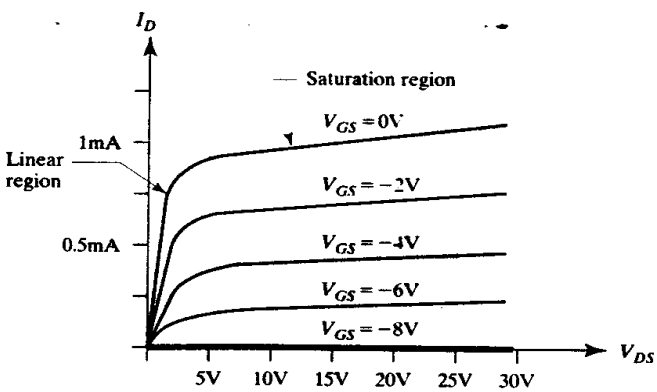


FIGURE 1.28 Typical I-V plot of an N-JFET transistor with $V_T = -8V$.

increases. Conduction ceases entirely when the gate voltage equals the *turnoff voltage* V_T. The turnoff voltage qualitatively corresponds to the threshold voltage of an MOS transistor. The comparison must not be taken too far, however, as the conduction equations of the two devices differ considerably.

The drain current curves of the N-JFET tilt slightly upward in saturation due to *channel length modulation*. This effect is analogous to that which occurs in MOS transistors. The pinched-off region of the JFET lengthens as the drain-to-source voltage increases. Any increase in the length of the pinched-off region produces a corresponding decrease in the length of the channel. The effect of channel length modulation is usually quite small because the channel length greatly exceeds the length of the pinched-off region.

The source and drain terminals of a JFET can often be interchanged without affecting the performance of the device. The JFET structure of Figure 1.27A is an example of such a *symmetric* device. More complex JFET structures sometimes exhibit differences in source and drain geometries that render them *asymmetric*.

Almost all JFET structures short the gate and backgate terminals. Consider the device of Figure 1.27A. The channel is bounded on the left by the source, on the right by the drain, on the top by the gate, and on the bottom by the backgate. The drawing does not show what bounds the channel on the front or the rear. In most cases, these sides of the channel are also bounded by reverse-biased junctions that are extensions of the gate-body and backgate-body junctions. This arrangement necessitates shorting the gate and backgate.

Figure 1.29 shows the conventional schematic symbols for N-channel and P-channel JFET transistors. The arrowhead on the gate lead shows the orientation of the PN junction between the gate and the body of the device. The symbol does not explicitly identify the source and drain terminals, but most circuit designers orient the devices so that the drain of an N-JFET and the source of a P-JFET lie on top.

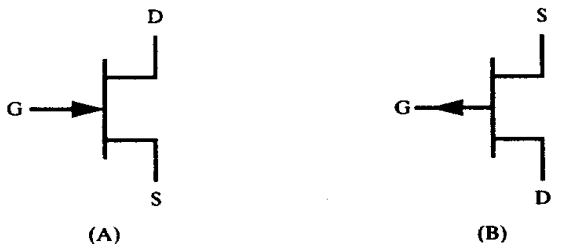


FIGURE 1.29 Symbols for an N-channel JFET (A) and a P-channel JFET (B).

1.6 SUMMARY

Device physics is a complex and ever-evolving science. Researchers constantly develop new devices and refine existing ones. Much of this ongoing research is highly theoretical and therefore lies beyond the scope of this introductory text. The functionality of most semiconductor devices can be satisfactorily explained using relatively simple and intuitive concepts.

This chapter emphasizes the role of majority and minority carrier conduction across PN junctions. If a junction is reverse-biased, then the majority carriers on either side of it are repelled and a depletion region forms. If the junction is forward-biased, then majority carriers diffuse across and recombine to create a net current flow across the PN junction. The PN junction diode employs this phenomenon to rectify signals.

When two junctions are placed in close proximity, carriers emitted by one junction can be collected by the other before they can recombine. The bipolar junction transistor (BJT) consists of just such a pair of closely spaced junctions. The voltage across the base-emitter junction of the BJT controls the current flowing from collector to emitter. If the transistor is properly designed, then a small base current can control a much larger collector current. The BJT therefore serves as an amplifier capable of transforming weak signals into much stronger ones. Thus, for example, a BJT can amplify a weak signal picked up by a radio receiver into a signal strong enough to drive a loudspeaker.

The metal-oxide-semiconductor (MOS) transistor relies upon electrical fields projected across a dielectric to modulate the conductivity of a semiconductor material. A suitable voltage placed upon the gate of a MOS transistor produces an electric field that attracts carriers up from the bulk silicon to form a conductive channel. The gate is insulated from the remainder of the transistor, so no gate current is required to maintain conduction. MOS circuitry can thus potentially operate at very low power levels.

The junction diode, the bipolar junction transistor, and the MOS transistor are the three most important semiconductor devices. Together with resistors and capacitors, they form the vast majority of the elements used in modern integrated circuits. The next chapter will examine how these devices are fabricated in a production environment.

1.7 EXERCISES

1.1. What are the relative proportions of aluminum, gallium, and arsenic atoms in intrinsic aluminum gallium arsenide?

1.2. A sample of pure silicon is doped with exactly 10^{16} atoms/cm^3 of boron and exactly 10^{16} atoms/cm^3 of phosphorus. Is the doped sample P-type or N-type?

1.3. The *instantaneous* velocity of carriers in silicon is almost unaffected by weak electric fields, yet the *average* velocity changes dramatically. Explain this observation in terms of drift and diffusion.

1.4. A layer of intrinsic silicon 1μ thick is sandwiched between layers of P-type and N-type silicon, both heavily doped. Draw a diagram illustrating the depletion regions that form in the resulting structure.

1.5. A certain process incorporates two different N+ diffusions that can be combined with a P– diffusion to produce Zener diodes. One of the resulting diodes has a breakdown voltage of 7V, while the other has a breakdown voltage of 10V. What causes the difference in breakdown voltages?

1.6. When the collector and emitter leads of an integrated NPN transistor are swapped, the transistor continues to function but exhibits a greatly reduced beta. There are several possible reasons for this behavior; explain at least one.

1.7. If a certain transistor has a beta of 60, and another transistor has a base twice as wide and half as heavily doped, then what is the approximate beta of the second transistor? What other electrical characteristics of the devices will vary, and how?

1.8. A certain MOS transistor has a threshold voltage of –1.5V. If a small amount of boron is added to the channel region, the threshold voltage shifts to –0.6V. Is the transistor PMOS or NMOS, and is it an enhancement or a depletion device?

1.9. If a depletion PMOS transistor has a threshold voltage of 0.5V when constructed using a 200Å oxide, will this threshold voltage increase or decrease if the oxide is thickened to 400Å?

1.10. A certain NMOS transistor has a threshold voltage of 0.5V; the gate-to-source voltage V_{GS} of the transistor is set to 2V, and the drain-to-source voltage V_{DS} is set to 4V.

What is the relative effect of doubling the gate-to-source voltage versus doubling the drain-to-source voltage, and why?

1.11. A certain silicon PN junction diode exhibits a forward voltage drop of 620mV when operated at a forward current of 25μA at a temperature of 25°C. What is the approximate forward drop of this diode at –40°C? At 125°C?

1.12. Two JFET transistors differ only in the separation between their gate and backgate; in one transistor these two regions are twice as far apart as in the other transistor. In what ways do the electrical properties of the two transistors differ?

2 Semiconductor Fabrication

Semiconductor devices have long been used in electronics. The first solid-state rectifiers were developed in the late nineteenth century. The galena crystal detector, invented in 1907, was widely used to construct crystal radio sets. By 1947, the physics of semiconductors was sufficiently understood to allow Bardeen and Brattain to construct the first bipolar junction transistor. In 1959, Kilby constructed the first integrated circuit, ushering in the era of modern semiconductor manufacture.

The impediments to manufacturing large quantities of reliable semiconductor devices were essentially technological, not scientific. The need for extraordinarily pure materials and precise dimensional control prevented early transistors and integrated circuits from reaching their full potential. The first devices were little more than laboratory curiosities. An entire new technology was required to mass produce them, and this technology is still rapidly evolving.

This chapter provides a brief overview of the process technologies currently used to manufacture integrated circuits. Chapter 3 then examines three representative process flows used for manufacturing specific types of analog integrated circuits.

2.1 SILICON MANUFACTURE

Integrated circuits are usually fabricated from *silicon*, a very common and widely distributed element. The mineral *quartz* consists entirely of silicon dioxide, also known as *silica*. Ordinary sand is chiefly composed of tiny grains of quartz and is therefore also mostly silica.

Despite the abundance of its compounds, elemental silicon does not occur naturally. The element can be artificially produced by heating silica and carbon in an electric furnace. The carbon unites with the oxygen contained in the silica, leaving more-or-less pure molten silicon. As this cools, numerous minute crystals form and grow together into a fine-grained gray solid. This form of silicon is said to be *polycrystalline* because it contains a multitude of crystals. Impurities and a disordered crystal structure make this *metallurgical-grade polysilicon* unsuited for semiconductor manufacture.

Metallurgical-grade silicon can be further refined to produce an extremely pure semiconductor-grade material. Purification begins with the conversion of the crude silicon into a volatile compound, usually trichlorosilane. After repeated distillation, the extremely pure trichlorosilane is reduced to elemental silicon using hydrogen gas. The final product is exceptionally pure, but still polycrystalline. Practical integrated circuits can only be fabricated from single-crystal material, so the next step consists of growing a suitable crystal.

2.1.1. Crystal Growth

The principles of crystal growing are both simple and familiar. Suppose a few crystals of sugar are added to a saturated solution that subsequently evaporates. The sugar crystals serve as seeds for the deposition of additional sugar molecules. Eventually the crystals grow to be very large. Crystal growth would occur even in the absence of a seed, but the product would consist of a welter of small intergrown crystals. The use of a seed allows the growth of larger, more perfect crystals by suppressing undesired nucleation sites.

In principle, silicon crystals can be grown in much the same manner as sugar crystals. In practice, no suitable solvent exists for silicon, and the crystals must be grown from the molten element at temperatures in excess of 1400°C. The resulting crystals are at least a meter in length and ten centimeters in diameter, and they must have a nearly perfect crystal structure if they are to be useful to the semiconductor industry. These requirements make the process technically challenging.

The usual method for growing semiconductor-grade silicon crystals is called the *Czochralski process*. This process, illustrated in Figure 2.1, uses a silica crucible charged with pieces of semi-grade polycrystalline silicon. An electric furnace raises the temperature of the crucible until all of the silicon melts. The temperature is then reduced slightly and a small seed crystal is lowered into the crucible. Controlled cooling of the melt causes layers of silicon atoms to deposit upon the seed crystal. The rod holding the seed slowly rises so that only the lower portion of the growing crystal remains in contact with the molten silicon. In this manner, a large silicon crystal can be pulled centimeter-by-centimeter from the melt. The shaft holding the crystal rotates slowly to ensure uniform growth. The high surface tension of molten silicon distorts the crystal into a cylindrical rod rather than the expected faceted prism.

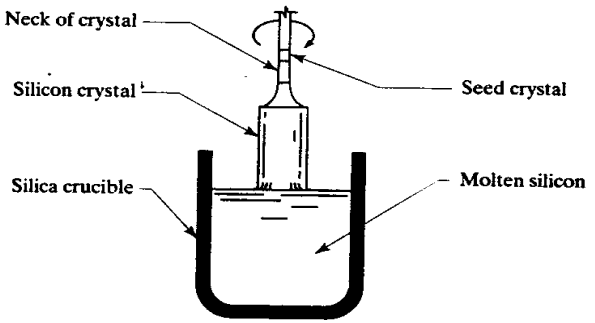


FIGURE 2.1 Czochralski process for growing silicon crystals.

The Czochralski process requires careful control to provide crystals of the desired purity and dimensions. Automated systems regulate the temperature of the melt and the rate of crystal growth. A small amount of doped polysilicon added to the melt sets the doping concentration in the crystal. In addition to the deliberately

introduced impurities, oxygen from the silica crucible and carbon from the heating elements dissolve in the molten silicon and become incorporated into the growing crystal. These impurities subtly influence the electrical properties of the resulting silicon. Once the crystal has reached its final dimensions, it is lifted from the melt and is allowed to slowly cool to room temperature. The resulting cylinder of monocrystalline silicon is called an *ingot.*

Since integrated circuits are formed upon the surface of a silicon crystal and penetrate this surface to no great depth, the ingot is customarily sliced into numerous thin circular sections called *wafers.* Each wafer yields hundreds or even thousands of integrated circuits. The larger the wafer, the more integrated circuits it holds and the greater the resulting economies of scale. Most modern processes employ either 150mm (6″) or 200mm (8″) wafers. A typical ingot measures between one and two meters in length and can provide hundreds of wafers.

2.1.2. Wafer Manufacturing

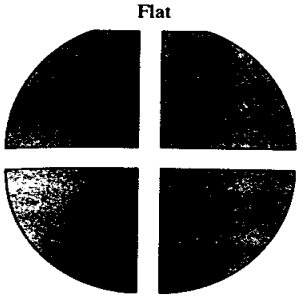
The manufacture of wafers consists of a series of mechanical processes. The two tapered ends of the ingot are sliced off and discarded. The remainder is then ground into a cylinder, the diameter of which determines the size of the resulting wafers. No visible indication of crystal orientation remains after grinding. The crystal orientation is experimentally determined and a flat stripe is ground along one side of the ingot. Each wafer cut from it will retain a facet, or *flat,* which unambiguously identifies its crystal orientation.

After grinding the flat, the manufacturer cuts the ingot into individual wafers using a diamond-tipped saw. In the process, about one-third of the precious silicon crystal is reduced to worthless dust. The surfaces of the resulting wafers bear scratches and pockmarks caused by the sawing process. Since the tiny dimensions of integrated circuits require extremely smooth surfaces, one side of each wafer must be polished. This process begins with mechanical abrasives and finishes with chemical milling. The resulting mirror-bright surface displays the dark gray color and characteristic near-metallic luster of silicon.

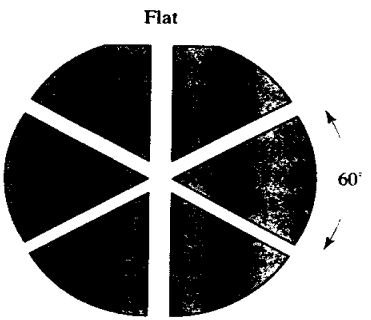
2.1.3. The Crystal Structure of Silicon

Each wafer constitutes a slice from a single silicon crystal. The underlying crystalline structure determines how the wafer splits when broken. Most crystals tend to part along *cleavage planes* where the interatomic bonding is weakest. For example, a diamond crystal can be cleaved by sharply striking it with a metal wedge. A properly oriented blow will split the diamond into two pieces, each of which displays a perfectly flat cleavage surface. If the blow is not properly oriented, then the diamond shatters. Silicon wafers also show characteristic cleavage patterns that can be demonstrated using a scrap wafer, a pad of note paper, and a wooden pencil. Place the wafer on the notepad, and place the pad in your lap. Take a wooden pencil and press down in the center of the wafer using the eraser. The wafer should split into either four or six regular wedge-shaped fragments, much like sections of a pie (Figure 2.2). The regularity of the cleavage pattern demonstrates that the wafer consists of monocrystalline silicon.

Figure 2.3 shows a small section of a silicon crystal drawn in three dimensions. Eighteen silicon atoms lie wholly or partially within the boundaries of an imaginary cube called a *unit cell.* Six of these occupy the centers of each of the six faces of the cube. Eight more atoms occupy the eight vertices of the cube. Two unit cells placed side-by-side share four vertex atoms and a single face-centered



(100) Wafer



(111) Wafer

FIGURE 2.2 Typical fracture patterns for (100) and (111) silicon wafers. Some wafers possess a second, smaller flat that denotes crystal orientation and doping. These *minor flats* have not been illustrated.

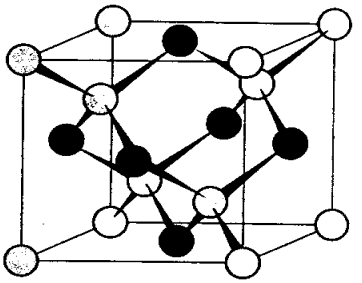


FIGURE 2.3 The diamond lattice unit cell displays a modified face-centered cubic structure. The face-centered atoms are shown in dark gray for emphasis.

atom. Additional unit cells can be placed on all sides to extend the crystal in all directions.

When the sawblade slices through a silicon ingot to form a wafer, the orientation of the resulting surface with respect to the unit cell determines many of the wafer's properties. A cut could, for example, slice across a face of the unit cell or diagonally through it. The pattern of atoms exposed by these two cuts differ, as do the electrical properties of devices formed into the respective surfaces. However, not all cuts made through a silicon crystal necessarily differ. Because the faces of a cube are indistinguishable from one another, a cut made across any face of the unit cell looks the same as cuts made across other faces. In other words, planes cut parallel to any face of a unit cube expose similar surfaces.

Because of the awkwardness of trying to describe various planes verbally, a trio of numbers called *Miller indices* are assigned to each possible plane passing through the crystal lattice (Appendix B). Figure 2.4 shows the two most important planar orientations. A plane parallel to a face of the cube is called a *(100) plane,* and a plane slicing diagonally through the unit cube to intersect three of its vertices is called a *(111) plane.* Silicon wafers are generally cut along either a (100) plane or a (111) plane. Although many other cuts exist, none of these have much commercial significance.

A trio of Miller indices enclosed in brackets denotes a direction perpendicular to the indicated crystal plane. For instance, a (100) plane has a [100] direction perpendicular to it and a (111) plane has a [111] direction perpendicular to it. Appendix B discusses how Miller indices are computed and explains the meaning of the different symbologies used to represent them.

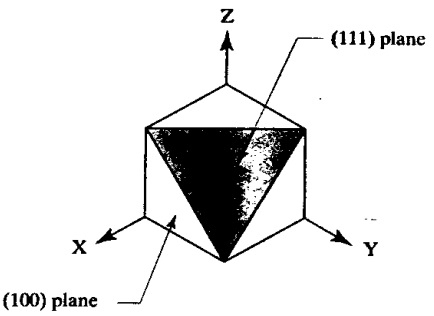


FIGURE 2.4 Identification of (100) and (111) planes of a cubic crystal.

2.2 PHOTOLITHOGRAPHY

The production of silicon wafers constitutes only the first step in the fabrication of integrated circuits. Many of the remaining steps deposit materials on the wafer or etch them away again. A variety of sophisticated deposition and etching techniques exist, but most of these are not *selective.* A nonselective, or *blanket,* process affects the entire surface of the wafer rather than just portions of it. The few processes that are selective are so slow or so expensive that they are useless for high-volume manufacturing. A technique called *photolithography* allows photographic reproduction of intricate patterns that can be used to selectively block depositions or etches. Integrated circuit fabrication makes extensive use of photolithography.

2.2.1. Photoresists

Photolithography begins with the application of a photosensitive emulsion called a *photoresist.* An image can be photographically transferred to the photoresist and a developer used to produce the desired masking pattern. The photoresist solution is usually *spun* onto the wafer. As shown in Figure 2.5, the wafer is mounted on a turntable spinning at several thousand revolutions per minute. A few drops of photoresist solution are allowed to fall onto the center of the spinning wafer, and centrifugal force spreads the liquid out across the surface. The photoresist solution adheres to the wafer and forms a uniform thin film. The excess solution flies off the edges of the spinning wafer. The film thins to its final thickness in a few seconds, the solvent rapidly evaporates, and a thin coating of photoresist remains on the wafer. This coating is baked to remove the last traces of solvent and to harden the photoresist to allow handling. Coated wafers are sensitive to certain wavelengths

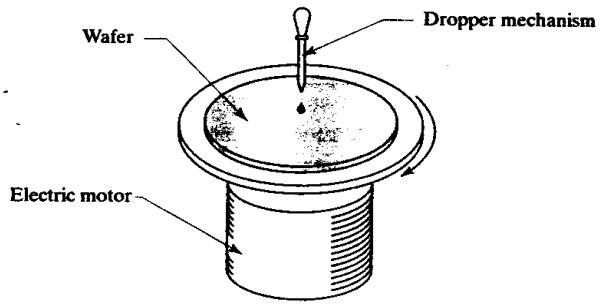


FIGURE 2.5 Application of photoresist solution to a wafer by spinning.

of light, particularly ultraviolet (UV) light. They remain relatively insensitive to other wavelengths, including those of red, orange, and yellow light. Most photolithography rooms therefore have special yellow lighting systems.

The two basic types of photoresists are distinguished by what chemical reactions occur during exposure. A *negative resist* polymerizes under UV light. The unexposed negative resist remains soluble in certain solvents, while the polymerized photoresist becomes insoluble. When the wafer is flooded with solvent, unexposed areas dissolve and exposed areas remain coated. A *positive resist,* on the other hand, chemically decomposes under UV light. These resists are normally insoluble in the developing solvent, but the exposed portions of the resist are chemically altered in order to become soluble. When the wafer is flooded with solvent, the exposed areas wash away while the unexposed areas remain coated. Negative resists tend to swell during development, so process engineers generally prefer to use positive resists.

2.2.2. Photomasks and Reticles

Modern photolithography depends upon a type of projection printing conceptually similar to that used to enlarge photographic negatives. Figure 2.6 shows a simplified illustration of the exposure process. A system of lenses collimates a powerful UV light source, and a plate called a *photomask* blocks the path of the resulting light beam. The UV light passes through the transparent portions of the photomask and through additional lenses that focus an image on the wafer. The apparatus in Figure 2.6 is called an *aligner* since it must ensure that the image of the mask aligns precisely with existing patterns on the wafer.

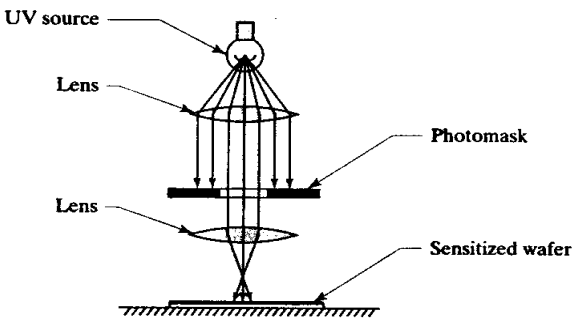


FIGURE 2.6 Simplified illustration of photomask exposure using an aligner.

The transparent plate used as the substrate of a photomask must be dimensionally stable or the pattern it projects will not align with those projected by previous masks. These plates most often consist of fused silica (often erroneously called quartz). After a thin layer of metal is applied to one surface of the plate, any one of various highly precise—but extremely slow and costly—methods are used to pattern the photomask. The image on the photomask is usually five or ten times the size of the image projected onto the wafer. Photographic reduction shrinks the size of any defects or irregularities in the photomask and therefore improves the quality of the final image. This type of enlarged photomask is called either a 5X or a 10X *reticle* depending on the degree of magnification employed.

A reticle can be used to directly pattern a wafer, but there are mechanical difficulties in doing so. The size of the photomask that an aligner can accept is limited by mechanical considerations, including the difficulty of constructing large lenses of the required accuracy. As a result, most commercial aligners accept a photomask about the same size as the wafer. A 5X reticle that could pattern an entire wafer in

one shot would be five times the size of the wafer, and would therefore not fit in the aligner. Practical 5X or 10X reticles are constructed to expose only a small rectangular portion of the final wafer pattern. The reticle must be stepped across the wafer and exposures made at many different positions in order to replicate the pattern across the entire wafer. This process is called *stepping*, and an aligner designed to step a reticle is called a *stepper*. Steppers are slower than ordinary aligners and are therefore more costly.

There is a faster method of exposing wafers that can be used for integrated circuits that do not require extremely fine feature sizes. The reticle can be stepped, not onto a sensitized wafer, but instead onto another photomask. This photomask now bears a 1X image of the desired pattern. The resulting photomask, called a *stepped working plate*, can expose an entire wafer in one shot. Stepped working plates make photolithography faster and cheaper, but the results are not as precise as directly stepping the reticle onto the wafer.

Even the tiniest dust speck is so large that it will block the transfer of a portion of the image and ruin at least one integrated circuit. Special air filtration techniques and protective garments are routinely used in wafer fabs, but some dust gets past all of these precautions. Photomasks are often equipped with *pellicles* on one or both sides to prevent dust from interfering with the exposure. Pellicles consist of thin transparent plastic films mounted on ring-shaped spacers that hold them slightly above the surface of the mask. Light passing through the plane of a pellicle is not in focus, so particles on the pellicle do not appear in the projected image. The pellicle also hermetically seals the surface of the mask and thereby protects it from dust.

2.2.3. Patterning

The exposed wafers are sprayed with a suitable developer, typically consisting of a mixture of organic solvents. The developer dissolves portions of the resist to uncover the surface of the wafer. A deposition or etch affects only these uncovered areas. Once the selective processing has been completed, the photoresist can be stripped away using solvents. Alternatively, the photoresist can be chemically destroyed by reactive ion etching in an oxygen ambient (Section 2.3.2). This procedure is called *ashing*.

Many important fabrication processes require masking layers that can withstand high temperatures. Since most practical photoresists are organic compounds, they are clearly unsuited to this task. Two common high-temperature masking materials are silicon dioxide and silicon nitride. These materials can be formed by the reaction of appropriate gases with the silicon surface. A photoresist can then be applied and patterned and an etching process used to open holes in the oxide or nitride film. Modern processing techniques make extensive use of oxide and nitride films for masking high-temperature depositions and diffusions.

2.3 OXIDE GROWTH AND REMOVAL

Silicon forms several oxides, the most important of which is *silicon dioxide* (SiO_2). This oxide possesses a number of desirable properties that together are so valuable that silicon has become the dominant semiconductor. Other semiconductors have better electrical properties, but only silicon forms a well-behaved oxide. Silicon dioxide can be grown on a silicon wafer by simply heating it in an oxidizing atmosphere. The resulting film is mechanically rugged and resists most common solvents, yet it readily dissolves in hydrofluoric acid. Oxide films are superb electrical insula-

tors and are useful not only for insulating metal conductor patterns but also for forming the dielectrics of capacitors and MOS transistors. Silicon dioxide is so important to silicon processing that it is universally known as *oxide.*

2.3.1. Oxide Growth and Deposition

The simplest method of producing an oxide layer consists of heating a silicon wafer in an oxidizing atmosphere. If pure dry oxygen is employed, then the resulting oxide film is called a *dry oxide.* Figure 2.7 shows a typical oxidation apparatus. The wafers are placed in a fused silica rack called a *wafer boat.* The wafer boat is slowly inserted into a fused silica tube wrapped in an electrical heating mantle. The temperature of the wafers gradually rises as the wafer boat moves into the middle of the heating zone. Oxygen gas blowing through the tube passes over the surface of each wafer. At elevated temperatures, oxygen molecules can actually diffuse through the oxide layer to reach the underlying silicon. There oxygen and silicon react, and the layer of oxide gradually grows thicker. The rate of oxygen diffusion slows as the oxide film thickens, so the growth rate decreases with time. As Table 2.1 indicates, high temperatures greatly accelerate oxide growth. Crystal orientation also affects oxidation rates, with (111) silicon oxidizing significantly faster than (100) silicon.[1] Once the oxide layer has reached the desired thickness (as gauged by time and temperature), the wafers are slowly withdrawn from the furnace.

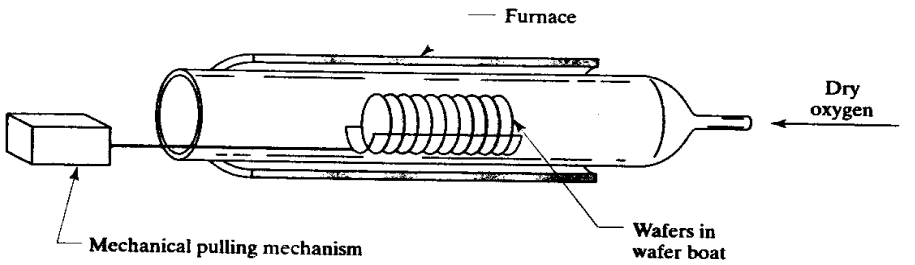


FIGURE 2.7 Simplified diagram of an oxidation furnace.

Ambient	800°C	900°C	1000°C	1100°C	1200°C
Dry O_2	30 hr	6 hr	1.7 hr	40 min	15 min
Wet O_2	1.7 hr	20 min	6 min		

TABLE 2.1 Times required to grow 0.1 μm of oxide on (111) silicon.[2]

Dry oxide grows very slowly, but it is of particularly high quality because relatively few defects exist at the oxide-silicon interface. These defects, or *surface states,* interfere with the proper operation of semiconductor devices, particularly MOS transistors. The density of surface states is measured by a parameter called the *surface state charge,* or Q_{ss}. Dry oxide films that are thermally grown on (100) silicon have especially low surface state charges and thus make ideal dielectrics for MOS transistors.

[1] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology* (Reading, MA: Addison-Wesley, 1994), p. 84ff.

[2] Calculated from R. P. Donovan, "Oxidation," in R. M. Burger and R. P. Donovan, eds., *Fundamentals of Silicon Integrated Device Technology* (Englewood Cliffs, NJ: Prentice-Hall, 1967), pp. 41, 49.

Wet oxides are formed in the same way as dry oxides, but steam is injected into the furnace tube to accelerate the oxidation. Water vapor moves rapidly through oxide films, but hydrogen atoms liberated by the decomposition of the water molecules produce imperfections that may degrade the oxide quality.[3] Wet oxidation is commonly used to grow a thick layer of *field oxide* where no active devices will be built. Dry oxidations conducted at higher-than-ambient pressures can also accelerate oxide growth rates.

Sometimes an oxide layer must be formed on a material other than silicon. For instance, oxide is frequently employed as an insulator between two layers of metallization. In such cases, some form of deposited oxide must be used rather than the grown oxides previously discussed. Deposited oxides can be produced by various reactions between gaseous silicon compounds and gaseous oxidizers. For example, silane gas and nitrous oxide react to form nitrogen gas, water vapor, and silicon dioxide. Deposited oxides tend to possess low densities and large numbers of defect sites, so they are not suitable for use as gate dielectrics for MOS transistors. Deposited oxides are still acceptable for use as insulating layers between multiple conductor layers, or as protective overcoats.

Oxide films are brightly colored due to *thin-film interference*. When light passes through a transparent film, destructive interference between transmitted and reflected wavefronts causes certain wavelengths of light to be selectively absorbed. Different thicknesses of films absorb different colors of light. Thin-film interference causes the iridescent colors seen in soap bubbles and films of oil on water. The same effect produces the vivid colors visible in microphotographs of integrated circuits. These colors are helpful in distinguishing various regions of an integrated circuit under a microscope or in a microphotograph. The approximate thickness of an oxide film can often be determined using a table of oxide colors.[4]

2.3.2. Oxide Removal

Figure 2.8 illustrates the procedure used to form a patterned oxide layer. The first step consists of growing a thin layer of oxide across the wafer. Next, photoresist is applied to the wafer by spinning. A subsequent oven bake drives off the final traces

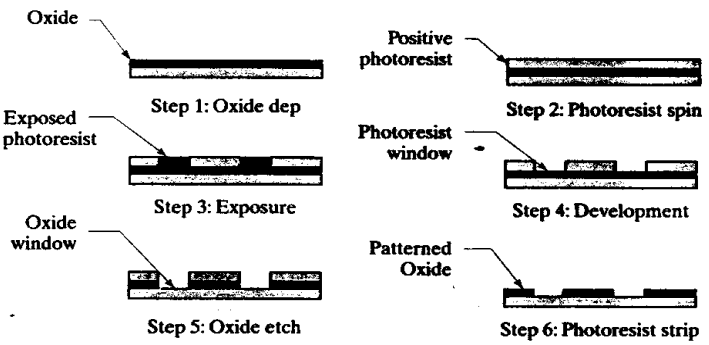


FIGURE 2.8 Steps in oxide growth and removal.

[3] Hydrogen incorporation due to wet oxidation conditions reduces the concentration of dangling bonds, but it increases the fixed oxide charge. The differences between wet and dry oxidation are therefore not as simplistic as the text may suggest.

[4] For a table, see W. A. Pliskin and E. E. Conrad, "Nondestructive Determination of Thickness and Refractive Index of Transparent Films," *IBM J. Research and Development*, Vol. 8, 1964, pp. 43–51.

of solvent and hardens the photoresist for handling. After photolithographic exposure, the wafer is developed by spraying it with a solvent that dissolves the exposed areas of photoresist to reveal the underlying oxide. The patterned photoresist serves as a masking material for an oxide etch. Having served its function, the photoresist is finally stripped away to leave the patterned oxide layer.

Oxide can be etched by either of two methods. *Wet etching* employs a liquid solution that dissolves the oxide, but not the photoresist or the underlying silicon. *Dry etching* uses a reactive plasma to perform the same function. Wet etches are simpler, but dry etches provide better linewidth control.

Most wet etches employ solutions of buffered hydrofluoric acid (HF). This highly corrosive substance readily dissolves silicon dioxide, but it does not attack either elemental silicon or organic photoresists. The etch process consists of immersing the wafers in a plastic tank containing the hydrofluoric acid solution for a specified length of time, followed by a thorough rinsing to remove all traces of the acid. Wet etches are *isotropic* because they proceed at the same rate laterally as well as vertically. The acid works its way under the edges of the photoresist to produce sloping sidewalls similar to those shown in Figure 2.9A. Since the etching must continue long enough to ensure that all openings have completely cleared, some degree of overetching inevitably occurs. The acid continues to erode the sidewalls as long as the wafer remains immersed. The extent of sidewall erosion varies depending upon etching conditions, oxide thickness, and other factors. Because of these variations, wet etching cannot provide the tight linewidth control required by modern semiconductor processes.

There are several types of dry etching processes.[5] One called *reactive ion etching* (RIE) employs plasma bombardment to erode the surface of the wafer. A silent electrical discharge passed through a low-pressure gas mixture forms highly energetic molecular fragments called *reactive ions*. The etching apparatus projects these ions downward onto the wafer at high velocities. Because the ions impact the wafer at a relatively steep angle, etching proceeds vertically at a much greater rate than laterally. The *anisotropic* nature of reactive ion etching allows the formation of nearly vertical sidewalls such as those shown in Figure 2.9B. Figure 2.10 shows a simplified diagram of a reactive ion etching apparatus.

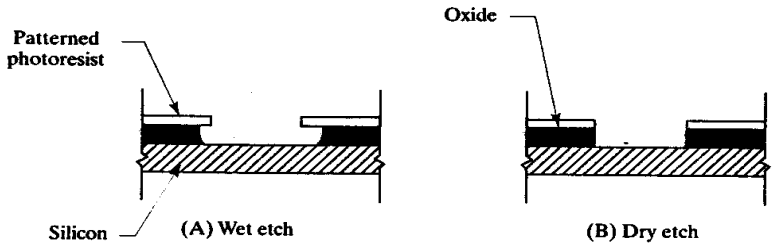


FIGURE 2.9 Comparison of isotropic wet etching (A) and anisotropic dry etching (B). Note the undercutting of the oxide caused by wet etching.

The etch gas employed in the RIE system generally consists of an organohalogen compound such as trichloroethane, perhaps mixed with an inert gas such as argon. The reactive ions formed from this mixture selectively attack silicon dioxide in preference to either photoresist or elemental silicon. Different mixtures of etch gases

[5] Reactive ion etching is actually only one of three forms of dry etching, the other two being plasma etching and chemical vapor etching. RIE is among the most useful because it produces highly anisotropic etching characteristics. See Runyan, *et al.*, pp. 269–272.

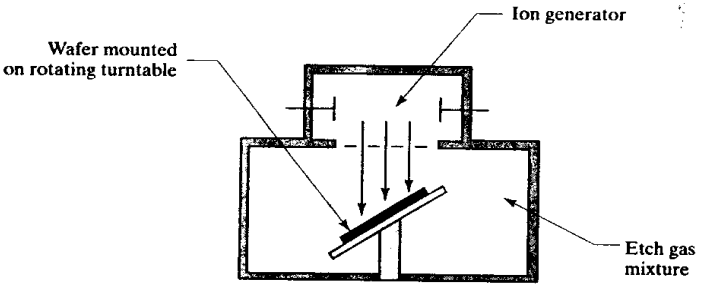


FIGURE 2.10 Simplified diagram of reactive ion etching apparatus.

have been developed that allow anisotropic etching of silicon nitride, elemental silicon, and other materials.

Modern processes rely on dry etching to obtain tight control of submicron geometries that cannot be fabricated in any other way. The increased packing density and higher performance of these structures more than compensate for the complexity and cost of dry etching.

2.3.3. Other Effects of Oxide Growth and Removal

During a typical processing sequence, the wafer is repeatedly oxidized and etched to form successive masking layers. These multiple masked oxidations cause the silicon surface to become highly nonplanar. The resulting surface irregularities are of great concern because modern fine-line photolithography has a very narrow depth of field. If the surface irregularities are too large, then it becomes impossible to focus the image of the photomask onto the resist.

Consider the wafer in Figure 2.11. A planar silicon surface has been oxidized, patterned, and etched to form a series of oxide openings (Figure 2.11A). Subsequent thermal oxidation of the patterned wafer results in the cross-section shown in Figure 2.11B. The opening that is left from the previous oxide removal initially oxidizes very rapidly, while the surfaces already coated with an oxide layer oxidize more slowly. The silicon surface erodes by about 45% of the oxide thickness grown.[6] The silicon under the previous oxide opening therefore recedes to a greater depth than the surrounding silicon surfaces. The thickness of oxide in the old opening will always be less than that of the surrounding surfaces since these already have some oxide on them when growth begins. The differences in oxide thickness and in the depths of the silicon surfaces combine to produce a characteristic surface discontinuity called an *oxide step.*

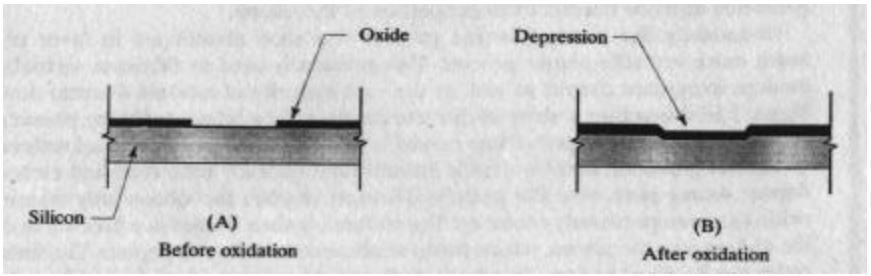


FIGURE 2.11 Effects of patterned oxidation on wafer topography.

[6] This value is the inverse of the *Pilling-Bedworth ratio,* which equals 2.2: G. E. Anner, *Planar Processing Primer* (New York: Van Nostrand Reinhold, 1990), p. 169.

The growth of a thermal oxide also affects the doping levels in the underlying silicon. If the dopant is more soluble in oxide than in silicon, during the course of the oxidation it will tend to migrate from the silicon into the oxide. The surface of the silicon thus becomes depleted of dopant. Boron is more soluble in oxide than in silicon, so it tends to segregate into the oxide. This effect is sometimes called *boron suckup.* Conversely, if the dopant dissolves more readily in silicon than in oxide, then the advancing oxide-silicon interface pushes the dopant ahead of it and causes a localized increase in doping levels near the surface. Phosphorus (like arsenic and antimony) segregates into the silicon, so it tends to accumulate at the surface as oxidation continues. This effect is sometimes called *phosphorus pileup* or *phosphorus plow.* The doping profiles of Figures 2.12A and 2.12B illustrate boron suckup and phosphorus plow, respectively. In both cases, the pre-oxidization doping profiles were constant and the varying dopant concentrations near the surface are solely due to segregation. The existence of these segregation mechanisms complicates the task of designing dopant profiles for integrated devices.

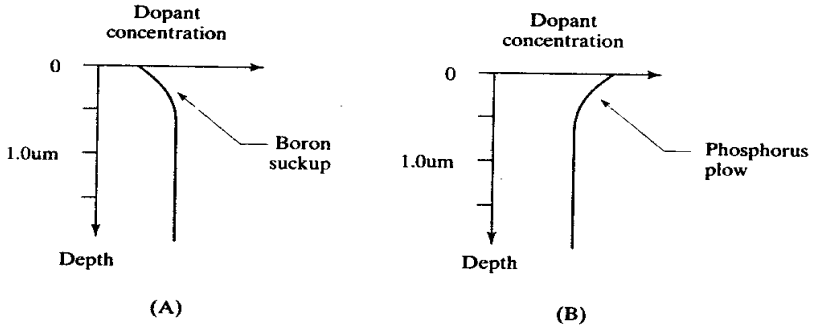


FIGURE 2.12 Oxide segregation mechanisms: (A) boron suckup and (B) phosphorus plow.[7]

The doping of silicon also affects the rate of oxide growth. A concentrated N+ diffusion tends to accelerate the growth of oxide near it by a process called *dopant-enhanced oxidation.* This occurs because the donors interfere with the bonding of atoms at the oxide interface, causing dislocations and other lattice defects. These defects catalyze oxidation and thus accelerate the growth of the overlying oxide. This effect can become quite significant when a heavily doped N+ deposition occurs early in the process, before the long thermal drives and oxidations. Figure 2.13 shows a wafer in which a long thermal oxidation has been

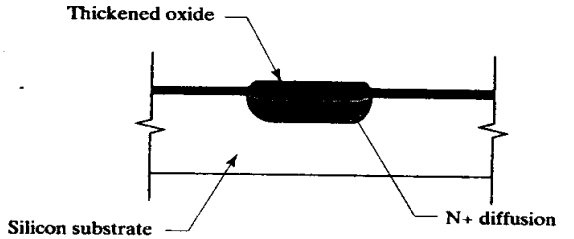


FIGURE 2.13 Effects of dopant-enhanced oxidation.

[7] A. S. Grove, O. Leistiko, and C. T. Sah, "Redistribution of Acceptor and Donor Impurities During Thermal Oxidation of Silicon," *J. Appl. Phys.*, Vol. 35, #9, 1964, pp. 2695–2701.

conducted after the deposition of an N+ region. The oxide over the N+ diffusion is actually thicker than the oxide over adjacent regions. Dopant-enhanced oxidation can be used to thicken the field oxide in order to reduce its capacitance per unit area. Thus, a capacitor formed over a deep-N+ diffusion will exhibit less parasitic capacitance between its bottom plate and the substrate than will a capacitor formed over lightly doped regions.

2.3.4. Local Oxidation of Silicon (LOCOS)

A technique called *local oxidation of silicon* (LOCOS) allows the selective growth of thick oxide layers.[8] The process begins with the growth of a thin *pad oxide* that protects the silicon surface from the mechanical stresses induced by subsequent processing (Figure 2.14). Chemical vapor deposition produces a nitride film on top of the pad oxide. This nitride is patterned to expose the regions to be selectively oxidized. The nitride blocks the diffusion of oxygen and water molecules, so oxidation only occurs under the nitride windows. Some oxidants diffuse a short distance under the edges of the nitride, producing a characteristic curved transition region called a *bird's beak*.[9] Once oxidation is complete, the nitride layer is stripped away to reveal the patterned oxide.

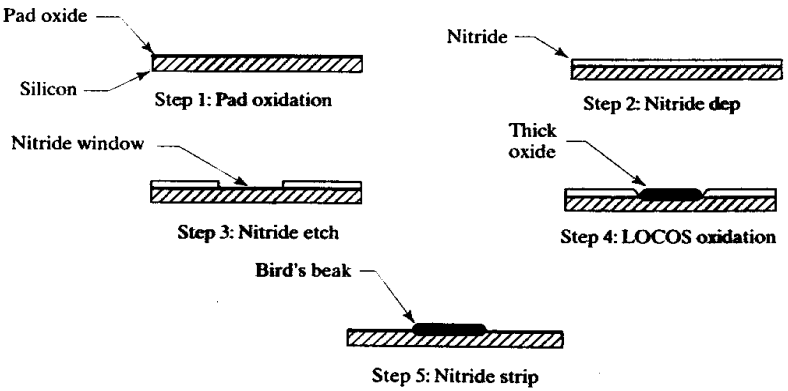


FIGURE 2.14 Local oxidation of silicon (LOCOS) process.

CMOS and BiCMOS processes employ LOCOS to grow a thick *field oxide* over electrically inactive regions of the wafer. The areas not covered by field oxide are called *moat* regions because they form shallow trenches in the topography of the wafer. A very thin, high-quality gate oxide subsequently grown in the moat regions forms the gate dielectric of the MOS transistors.

A mechanism called the *Kooi effect* complicates the growth of gate oxide.[10] The water vapor typically used to accelerate LOCOS oxidation also attacks the surface of the nitride film to produce ammonia, some of which migrates beneath the pad oxide near the edges of the nitride window. There it reacts with the underlying silicon to form silicon nitride again (Figure 2.15). Since these nitride deposits lie

---

[8] "LOCOS: A New I.C. Technology," *Microelectronics and Reliability*, Vol. 10, 1971, pp. 471–472.

[9] E. Bassous, H. N. Yu, and V. Maniscalco, "Topology of Silicon Structures with Recessed SiO_2," *J. Electrochem. Soc.*, Vol. 123, #11, 1976, pp. 1729–1737.

[10] E. Kooi, J. G. van Lierop, and J. A. Appels, "Formation of Silicon Nitride at a Si-SiO_2 Interface during Local Oxidation of Silicon and during Heat-Treatment of Oxidized Silicon in NH_3 Gas," *J. Electrochem. Soc.*, Vol. 123, #7, 1976, pp. 1117–1120.

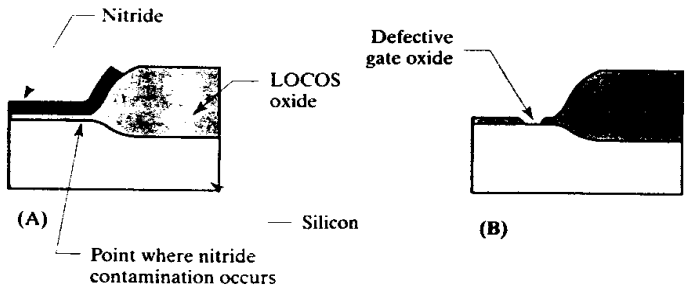


FIGURE 2.15 The Kooi effect is caused by nitride that grows under the bird's beak (A), preventing formation of gate oxide during subsequent oxidation (B).

beneath the pad oxide, they remain even after the LOCOS nitride is stripped. Removing the pad oxide prior to growing the gate oxide does not eliminate these deposits because this etch is selective to oxide, not to nitride. During gate oxidation, the nitride residues act as an unintentional LOCOS mask that retards oxide growth around the edges of the moat region. The gate oxide at these points may not be sufficiently thick to withstand the full operating voltage. The Kooi effect can be circumvented by first growing a thin oxide layer and then stripping it away. Because silicon nitride slowly oxidizes, this *dummy gate oxidation* removes the nitride residues and improves the integrity of the true gate oxide grown immediately afterward.

2.4 DIFFUSION AND ION IMPLANTATION

Discrete diodes and transistors can be fabricated by forming junctions into a silicon ingot during crystal growth. Suppose that the silicon ingot begins as a P-type crystal. After a short period of growth, the melt is counterdoped by the addition of a controlled amount of phosphorus. Continued crystal growth will now produce a PN junction embedded in the ingot. Successive counterdopings can produce multiple junctions in the crystal, allowing the fabrication of *grown-junction* transistors. Integrated circuits cannot be grown because there is no way to produce differently doped regions in different portions of the wafer. Even the manufacture of simple grown-junction transistors presents a challenge, because the thickness and planarity of grown junctions are difficult to control. Each counterdoping also raises the total dopant concentration. Some properties of silicon (such as minority carrier lifetime) depend upon the total concentration of doping atoms, not just upon the excess of one dopant species over the other. The repeated counterdopings therefore progressively degrade the electrical properties of the silicon.

Historically, the grown junction process was soon abandoned in favor of the much more versatile *planar process*. This process is used to fabricate virtually all modern integrated circuits as well as the vast majority of modern discrete devices. Figure 2.16 shows how a wafer of discrete diodes can be fabricated using planar processing. A uniformly doped silicon crystal is first sliced to form individual wafers. An oxide film grown on these wafers is photolithographically patterned and etched. A dopant source spun onto the patterned wafers touches the silicon only where the oxide has been previously removed. The wafers are then heated in a furnace to drive the dopant into the silicon, which forms shallow counterdoped regions. The finished wafer can be diced to form hundreds or thousands of individual diodes. The planar process does not require multiple counterdopings of the silicon ingot, thereby allowing more precise control of junction depths and dopant distributions.

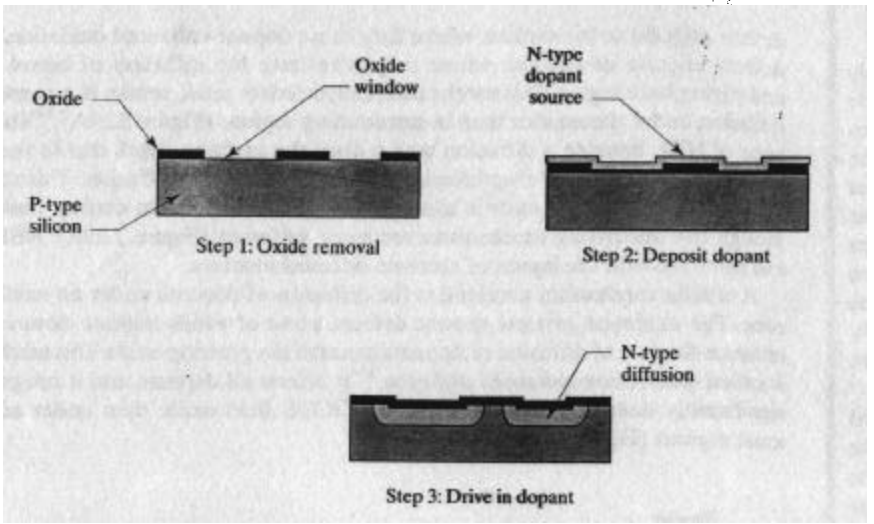


FIGURE 2.16 Formation of diffused PN-junction diodes using the planar process.

2.4.1. Diffusion

Dopant atoms can move through the silicon lattice by thermal diffusion in much the same way as carriers move by diffusion (Section 1.1.3). The heavier dopant atoms are more tightly bound to the crystal lattice, so temperatures of 800°C to 1250°C are required to obtain reasonable diffusion rates. Once the dopants have been driven to the desired junction depth, the wafer is cooled and the dopant atoms become immobilized within the lattice. A doped region formed in this manner is called a *diffusion.*

The usual process for creating a diffusion consists of two steps: an initial *deposition* (or *predeposition*) and a subsequent *drive* (or *drive-in*). Deposition consists of heating the wafer in contact with an external source of dopant atoms. Some of these diffuse from the source into the surface of the silicon wafer to form a shallow heavily doped region. The external dopant source is then removed and the wafer is heated to a higher temperature for a prolonged period of time. The dopants introduced during deposition are now driven down to form a much deeper and less concentrated diffusion. If a very heavily doped junction is required, then it is usually unnecessary to strip the dopant source from the wafer, and the deposition and subsequent drive can be conducted as a single operation.

Four dopants find widespread use in silicon processing: *boron, phosphorus, arsenic,* and *antimony.*[11] Only boron is an acceptor; the other three are all donors. Boron and phosphorus diffuse relatively rapidly, while arsenic and antimony diffuse much more slowly (Table 2.2). Arsenic and antimony are used where slow rates of

TABLE 2.2 Representative junction depths, in microns (10^{20} atoms/cm^3 source, 10^{16} atoms/cm^3 background, 15 min deposition, 1 hr drive).[12]

Dopant	950°C	1000°C	1100°C	1200°C
Boron	0.9	1.5	3.6	7.3
Phosphorus		0.5	1.6	4.6
Antimony			0.8	2.1
Arsenic			0.7	2.0

[11] These dopants were chosen because they readily ionize and because they are sufficiently soluble in silicon to form heavily doped diffusions. See F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell Syst. Tech. J.*, Vol. 39, #1, 1960, pp. 205–233.

[12] Calculated using diffusivities from R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons, 1986), p. 85.

diffusion are advantageous—for example, when very shallow junctions are desired. Even boron and phosphorus do not diffuse appreciably at temperatures below 800°C, necessitating the use of special high-temperature diffusion furnaces.

Figure 2.17 shows a simplified diagram of a typical apparatus for conducting a phosphorus diffusion. A long fused silica tube passes through an electric furnace that is constructed to produce a very stable heating zone in the middle of the tube. After the wafers are loaded into a wafer boat, they are slowly pushed into the furnace by means of a mechanical arrangement that controls the insertion rate. Dry oxygen is blown through a flask containing liquid phosphorus oxychloride ($POCl_3$, often called "pockle"). A small amount of $POCl_3$ evaporates and is carried by the gas stream over the wafers. Phosphorus atoms released by the decomposition of the $POCl_3$ diffuse into the oxide film, forming a doped oxide that acts as a deposition source. When enough time has passed to deposit sufficient dopant in the silicon, the wafers are removed from the furnace and the doped oxide is stripped away (a process called *deglazing*). The wafers are then reloaded into another furnace, where they are heated to drive the phosphorus down to form the desired diffusion. If a very concentrated phosphorus diffusion is desired, then the wafers need not be removed for deglazing prior to the drive. With suitable modifications to the dopant source, this apparatus can diffuse any of the four common dopants.

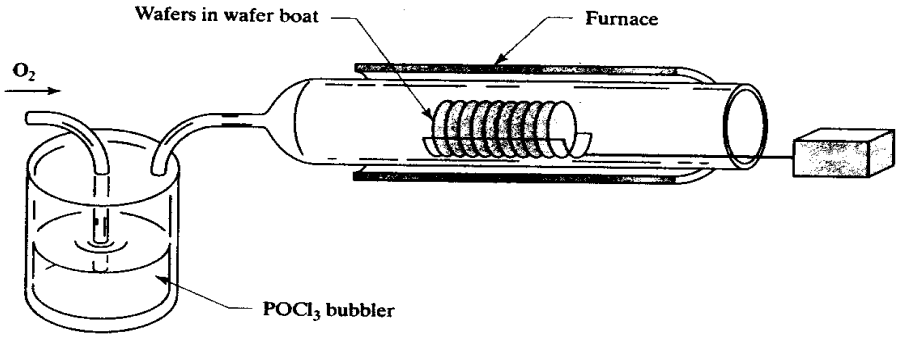


FIGURE 2.17 Simplified diagram of a phosphorus diffusion furnace using a $POCl_3$ source.

Many alternate deposition sources have been developed. A gaseous dopant such as diborane (for boron) or phosphine (for phosphorus) can be injected directly into the carrier gas stream. Thin disks of boron nitride placed between silicon wafers can serve as a solid deposition source for boron. In a high-temperature oxidizing atmosphere, a little boron trioxide outgases from these disks to the adjacent wafers. Various proprietary *spin-on glasses* are also sold as dopant sources. These consist of doped oxide dispersed in a volatile solvent. After the solution is spun onto a wafer, a brief bake drives out the solvent and leaves a doped oxide layer on the wafer. This so-called *glass* then serves as a dopant source for the subsequent diffusion.

None of these deposition schemes are particularly well controlled. Even with gaseous sources (which can be precisely metered) nonuniform gas flow around the wafer inevitably produces doping variations. For less-demanding processes such as standard bipolar, any of these schemes can give adequate results. Modern CMOS and BiCMOS processes require more accurate control of doping levels and junction depths than conventional deposition techniques can achieve. Ion implantation can provide the necessary accuracy at the expense of much more complex and costly apparatus.

2.4.2. Other Effects of Diffusion

The diffusion process suffers from a number of limitations. Diffusions can only be performed from the surface of the wafer, limiting the geometries that can be fabricated. Dopants diffuse unevenly, so the resulting diffusions do not have constant doping profiles. Subsequent high-temperature process steps continue the drive of previously deposited dopants, so junctions formed early in the process are driven substantially deeper during later processing. Dopants out-diffuse under the edges of the oxide windows, spreading the diffusion pattern. Diffusions interact with oxidizations due to segregation mechanisms, resulting in depletion or enhancement of surface doping levels. Diffusions even interact with one another since the presence of one doping species alters the diffusion rates of others. These and other complications make the diffusion process far more complex than it might at first appear.

Diffusion can produce only relatively shallow junctions. Practical drive times and temperatures limit junction depths to about fifteen microns. Most diffusions will be much shallower. Since diffusions are typically patterned using an oxide mask, the cross section of a diffusion generally resembles that shown in Figure 2.18. The dopant diffuses out in all directions at roughly the same rate. The junction moves laterally under the edges of the oxide window a distance equal to about 80% of the junction depth.[13] This lateral movement, known as *outdiffusion*, causes the final size of the diffused region to exceed the drawn dimensions of the oxide window. Outdiffusion is not visible under the microscope since the changes in oxide color caused by thin film interference correspond to the locations of oxide removals and not to the positions of the final junctions.

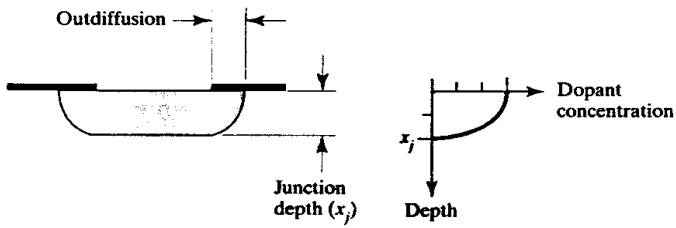


FIGURE 2.18 Cross section and doping profile of a typical planar diffusion.

The doping level of a diffusion varies as a function of depth. Neglecting segregation mechanisms, dopant concentrations are highest at the surface and gradually lessen with depth. The resulting ***doping profile*** can be theoretically predicted and experimentally measured. Figure 2.18 shows the theoretical doping profile for a point in the center of the oxide window. This profile assumes that oxide segregation remains negligible, which is not always the case. Boron suckup may substantially reduce the surface doping of a P-type diffusion and can cause a lightly doped diffusion to invert to become N-type. Phosphorus pileup will not cause surface inversion, but it still affects surface doping levels.

As mentioned above, the rate of diffusion can be altered by the presence of other doping species. Consider an NPN transistor with a heavily doped phosphorus emitter diffused into a lightly doped boron base. The presence of high concentrations of donors within the emitter causes lattice strains that spawn defects. Some of these

[13] See D. P. Kennedy and R. R. O'Brien, "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar p-n Junction," *IBM J. of Research and Development*, Vol. 9, 1965, pp. 179–186.

defects migrate to the surface, where they cause dopant-enhanced oxidation. Other defects migrate downward, where they accelerate the diffusion of boron in the underlying base region. This mechanism, called *emitter push*, results in a deeper base diffusion under the emitter than in surrounding regions (Figure 2.19A).[14] The presence of NBL beneath a diffusion may reduce the junction depth due to the intersection of the tail of the updiffusing NBL with the base diffusion. This effect is sometimes called *NBL push* in analogy with the better-known emitter push, even though the underlying mechanisms are quite different (Figure 2.19B). NBL push can interfere with the layout of accurate diffused resistors.

A similar mechanism accelerates the diffusion of dopants under an oxidization zone. The oxidation process spawns defects, some of which migrate downward to enhance the rate of diffusion of dopants beneath the growing oxide. This mechanism is called *oxidization-enhanced diffusion*.[15] It affects all dopants, and it can produce significantly deeper diffusions under a LOCOS field oxide than under adjacent moat regions (Figure 2.19C).

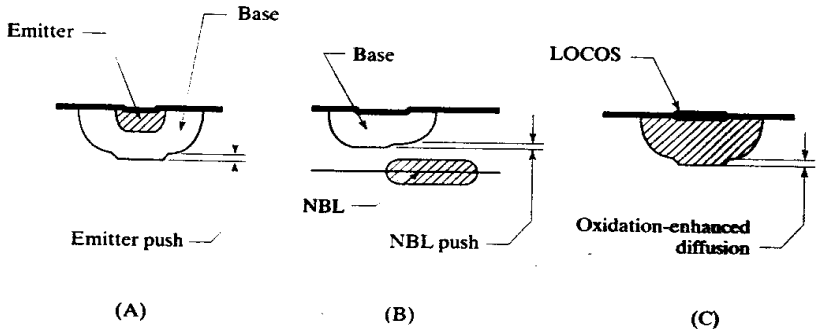


FIGURE 2.19 Mechanisms that can alter diffusion rates include emitter push (A), NBL push (B), and oxidation-enhanced diffusion (C).

Even the most sophisticated computer programs cannot always predict actual doping profiles and junction depths because of the many interactions that occur. Process engineers must experiment carefully to find the proper recipe for manufacturing a given combination of devices on a wafer. The more complicated the process, the more complex these interactions become and the more difficult it is to find a suitable recipe. Since process design takes so much time and effort, most companies use only a few processes to manufacture all of their products. The difficulty of incorporating new process steps into an existing recipe also explains the reluctance of process engineers to modify their processes.

2.4.3. Ion Implantation

Due to the limitations of conventional diffusion techniques, modern processes make extensive use of *ion implantation*. An ion implanter is essentially a specialized particle accelerator used to accelerate dopant atoms so that they can penetrate the silicon crystal to a depth of several microns.[16] Ion implantation does not require high

[14] A. F. W. Willoughby, "Interactions between Sequential Dopant Diffusions in Silicon—A Review," *J. Phys. D: Appl. Phys.*, Vol. 10, 1977, pp. 455–480.

[15] K. Taniguchi, K. Kurosawa, and M. Kashiwagi, "Oxidation Enhanced Diffusion of Boron and Phosphorus in (100) Silicon," *J. Electrochem. Soc.*, Vol. 127, #10, 1980, p. 2243–2248.

[16] The depth of the implant depends on implant energy. The implants discussed in this section involve energies of no more than several hundred keV. Some modern CMOS processes now employ multi-MeV implants to achieve significantly deeper profiles (5–10μm).

temperatures, so a layer of patterned photoresist can serve as a mask against the implanted dopants. Implantation also allows better control of dopant concentrations and profiles than conventional deposition and diffusion. However, large implant doses require correspondingly long implant times. Ion implanters are also complex and costly devices. Many processes use a combination of diffusions and implantations to reduce overall costs.

Figure 2.20 shows a simplified diagram of an ion implanter. An ion source provides a stream of ionized dopant atoms that are accelerated by the electric field of a miniature linear accelerator. A magnetic analyzer selects the desired species of ion, and a pair of deflection plates scans the resulting ion beam across the wafer. A high vacuum must be maintained throughout the system, so the entire apparatus is enclosed in a steel housing.

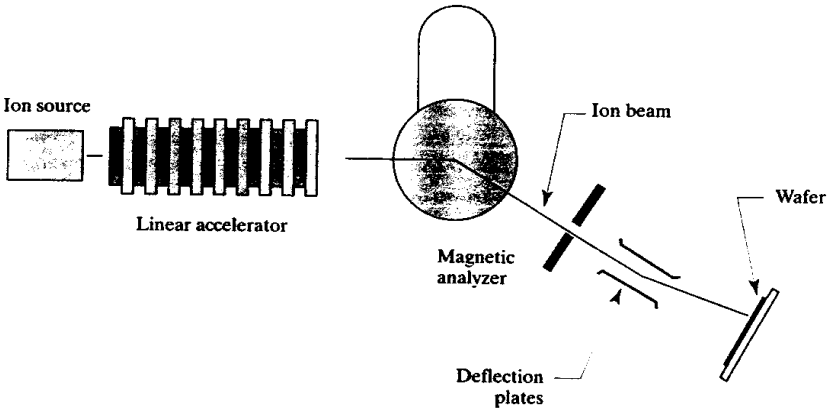


FIGURE 2.20 Simplified diagram of an ion implanter.[17]

Once the ions enter the silicon lattice, they immediately begin to decelerate due to collisions with surrounding atoms. Each collision transfers momentum from a moving ion to a stationary atom. The ion beam rapidly spreads as it sheds energy, causing the implant to spread out (*straggle*) in a manner reminiscent of outdiffusion. Atoms are also knocked out of the lattice by the collisions, causing extensive lattice damage that must be repaired by *annealing* the wafer at moderate temperatures (800°C to 900°C) for a few minutes. The silicon atoms become mobile and the intact silicon crystal structure around the edges of the implant zone serves as a seed for crystal growth. Damage progressively anneals out from the sides of the implant zone toward the center. Dopants added by ion implantation will redistribute by thermal diffusion if the wafer is subsequently heated to a sufficiently high temperature. Therefore, a deep lightly doped diffusion can be created by first implanting the required dopants and subsequently driving them down to the desired junction depth.

The dopant concentration provided by ion implantation is directly proportional to the *implant dose*, which equals the product of ion beam current and time. The dose can be precisely monitored and controlled, which allows for much better repeatability than conventional deposition techniques do. The doping profile is determined by the energy imparted to individual ions, a quantity called the *implant energy*. Low-energy implants are very shallow, while high-energy implants actually place most of the dopant atoms beneath the surface of the silicon. Ion implantation

[17] The scheme shown is but one of several; see Anner, p. 313ff.

can be used to counterdope a subsurface region to form a *buried layer.* Because of practical limitations on implant energy, these buried layers are usually quite shallow.

Ion implantation is somewhat anisotropic. The edges of an implant, especially a shallow low-energy one, do not spread as much as those produced by thermal diffusion. This aids in the manufacture of *self-aligned* structures that greatly improve the performance of MOS transistors. Figure 2.21 illustrates the creation of self-aligned MOS source/drain regions by ion implantation. A layer of polysilicon has been deposited and patterned on top of a thin gate oxide. The polysilicon not only forms the gate electrodes for MOS transistors but also simultaneously serves as an implant mask. The polysilicon blocks the implant from the region beneath the gate electrode, forming precisely aligned source and drain regions. The alignment of the source and drain with the gate is limited only by the small amount of straggle caused by the spreading of the ion beam. If self-aligned implants were not used, then photolithographic misalignments would occur between the gate and the source/drain diffusions, and the resulting overlap capacitances would substantially reduce the switching speed of the MOS transistors.

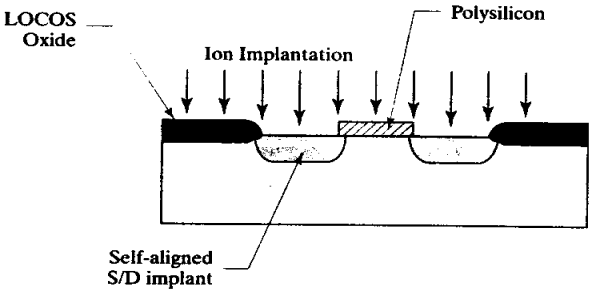


FIGURE 2.21 Self-aligned source and drain regions formed by ion implantation.

When the silicon lattice is viewed from certain angles, interstices between columns of silicon atoms, called *channels,* become visible. These disappear from view when the crystal is turned slightly. Channels are visible in both the (100) and the (111) silicon surfaces when these are viewed perpendicularly. If the ion beam were to impinge perpendicularly upon a (100) or a (111) surface, then ions could move deep into the crystal before scattering would commence. The final dopant distribution would depend critically upon the angle of incidence of the ion beam. To avoid this difficulty, most implanters project the ion beam onto the wafer at an angle of about 7°.

2.5 SILICON DEPOSITION

Films of pure or doped silicon can be chemically grown on the surface of a wafer. The nature of the underlying surface determines whether the resulting film will be monocrystalline or polycrystalline. If the surface consists of exposed monocrystalline silicon, then this serves as a seed for crystal growth and the deposited film will also be monocrystalline. If the deposition is conducted on top of an oxide or nitride film, then no underlying crystalline lattice will exist to serve as a seed for crystal nucleation, and the deposited silicon will form a fine-grained aggregate of polycrystalline silicon (*poly*). Modern integrated circuits make extensive use of both monocrystalline and polycrystalline deposited silicon films.

2.5.1. Epitaxy

The growth of a single-crystal semiconductor film upon a suitable crystalline substrate is known as *epitaxy*. The substrate normally consists of a crystal of the same material as the semiconductor that is to be deposited, but this need not always be the case. High-quality monocrystalline silicon films have been grown on wafers of synthetic sapphire or spinel, as these materials possess a crystal structure that is enough like silicon to allow crystal nucleation. The cost of synthetic sapphire or spinel wafers so greatly exceeds the cost of similar-sized silicon wafers that the vast majority of epitaxial depositions consist of silicon films grown on silicon substrates.

There are several different methods of growing epitaxial (*epi*) layers. One relatively crude method consists of pouring molten semiconductor material on top of the substrate, allowing it to crystallize for a short period of time, and then wiping the excess liquid off. The wafer surface can then be reground and polished to form an epitaxial layer. Obvious drawbacks to this *liquid-phase epitaxy* include the high cost of regrinding the wafer and the difficulty of producing a precisely controlled epi thickness.

Most modern epitaxial depositions use *low pressure chemical vapor deposited* (LPCVD) epitaxy. Figure 2.22 shows a simplified diagram of an early type of LPCVD epi reactor. The wafers are mounted on an inductively heated carrier block, and a mixture of dichlorosilane and hydrogen passes over them. These gases react at the surface of the wafers to form a slow-growing layer of monocrystalline silicon. The rate of growth can be controlled by adjusting the temperature, pressure, and gas mixture used in the reactor. No polishing is required to render the epitaxial surface suitable for further processing, as vapor-phase epitaxy faithfully reproduces the topography of the underlying surface. The epitaxial film can also be doped *in situ* by adding small amounts of gaseous dopants such as phosphine or diborane to the gas stream.

There are several benefits of growing an epitaxial layer on the starting wafer. For one, the epi layer need not have the same doping polarity as the underlying wafer. For example, an N– epitaxial layer can be grown on a P– substrate—an arrangement commonly employed for standard bipolar processes. Multiple epitaxial layers can be grown in succession and the resulting stack can be used to form transistors or other devices. The potential of epitaxy is limited chiefly by the slow rate of epi growth and by the expense and complexity of the required equipment, which are much greater than Figure 2.22 suggests.

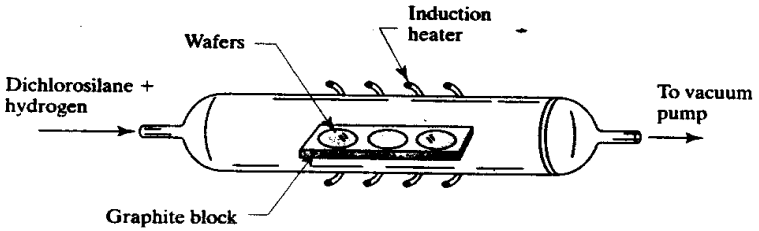


FIGURE 2.22 Simplified diagram of an epi reactor.[18]

[18] The horizontal tube reactor shown here has long been obsolete; see C. W. Pearce, "Epitaxy," in S. M. Sze, ed., *VLSI Technology* (New York: McGraw-Hill, 1983), pp. 61–65.

Epitaxy also allows the formation of *buried layers*. An N+ buried layer constitutes a key step in most bipolar processes since it allows the construction of vertical NPN transistors with low collector resistances. Figure 2.23 depicts the growth of such an N-buried layer (NBL). Arsenic and antimony are the preferred dopants for forming an NBL because their slow diffusion rates minimize the outdiffusion of the buried layer during subsequent high-temperature processing. Antimony is often chosen instead of arsenic because it exhibits less tendency to spread laterally during epitaxy (an effect called *lateral autodoping*).[19] Buried layer fabrication begins with a lightly doped P-type wafer. This wafer is oxidized, and windows are patterned in the resulting oxide layer. Either arsenic or antimony is implanted through the windows, and the wafer is briefly annealed to eliminate the resulting implant damage. Thermal oxidation occurs during this anneal, and discontinuities form around the edges of the oxide windows. Next, all oxide is stripped from the wafer, and an N– epitaxial layer is deposited. The resulting structure consists of patterned N+ regions buried under an epitaxial layer.

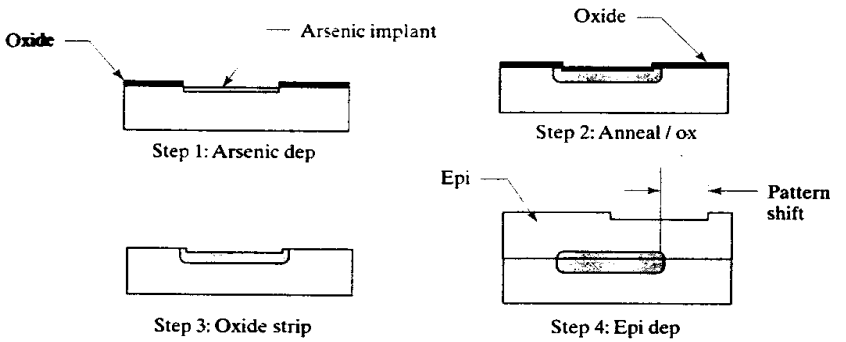


FIGURE 2.23 Formation of an N-buried layer (NBL), showing pattern shift.

As mentioned previously, oxidization during the anneal of the NBL causes slight surface discontinuities to form around the edges of the oxide window. The epitaxial layer faithfully reproduces these discontinuities in the final surface of the wafer. Under a microscope, the resulting step forms a faintly visible outline called the *NBL shadow*. Subsequent photomasks are aligned to this discontinuity. An alternative alignment method uses infrared light to image the NBL doping through the overlying silicon, but this requires more complicated equipment.

The accretion of silicon atoms at the edge of the NBL shadow during epitaxy displaces it laterally, an effect called *pattern shift* (Figure 2.23).[20] The magnitude of shift depends on many factors, including temperature, pressure, gas composition, substrate orientation, and tilt (See Section 7.2.3). When other layers are aligned to the NBL shadow, these must be offset to compensate for pattern shift.

---

[19] M. W. M. Graef, B. J. H. Leunissen, and H. H. C. de Moor, "Antimony, Arsenic, Phosphorus, and Boron Autodoping in Silicon Epitaxy," *J. Electrochem. Soc.*, Vol. 132, #8, 1985, pp. 1942–1954.

[20] M. R. Boydston, G. A. Gruber, and D. C. Gupta, "Effects of Processing Parameters on Shallow Surface Depressions During Silicon Epitaxial Deposition," in *Silicon Processing*, American Society for Testing and Materials STP 804, 1983, pp. 174–189.

2.5.2. Polysilicon Deposition

If silicon is deposited on an amorphous material, then no underlying lattice exists to align crystal growth. The resulting silicon film consists of an aggregate of small intergrown crystals. This *poly* film has a granular structure with a grain size dependent upon deposition conditions and subsequent heat treatment. The grain boundaries of the poly represent lattice defects, which can provide sneak paths for leakage currents. Therefore, PN junctions are not normally fabricated from poly. Polysilicon is often used to construct the gate electrodes of self-aligned MOS transistors because, unlike aluminum, it can withstand the high temperatures required to anneal the source/drain implants. In addition, the use of poly has led to better control of MOS threshold voltages due to the ability of phosphorus-doped polysilicon to immobilize ionic contaminants (Section 4.2.2). Suitably doped poly can be used to fabricate very narrow resistors that exhibit fewer parasitics than diffused devices. Heavily doped polysilicon can also be used as an additional metallization layer for signals that can tolerate the insertion of considerable resistance in the signal path.

A patterned poly layer is produced by first depositing polysilicon across the wafer using an apparatus similar to that employed for epitaxy (see Figure 2.22). The wafer is then coated with photoresist, patterned, and etched to selectively remove the polysilicon. Modern processes usually employ dry etching rather than wet etching because of the importance of precisely controlled gate dimensions.

2.6 METALLIZATION

The active components of an integrated circuit consist of diffusions, ion implantations, and epitaxial layers grown in or on a silicon substrate. When this processing is complete, the resulting components are connected to form the integrated circuit, using one or more layers of patterned wiring. This wiring consists of layers of metal and polysilicon separated by insulating material, usually deposited oxide. These same materials can also be used to construct passive components such as resistors and capacitors.

Figure 2.24 illustrates the formation of a typical *single-level-metal* (SLM) interconnection system. After the final implantations and diffusions, a layer of oxide is grown or deposited over the entire wafer, and selected areas are patterned and etched to create oxide windows exposing the silicon. These windows will form *contacts* between the metallization and the underlying silicon. Once these contacts have been opened, a thin metallic film can be deposited and etched to form the interconnection pattern.

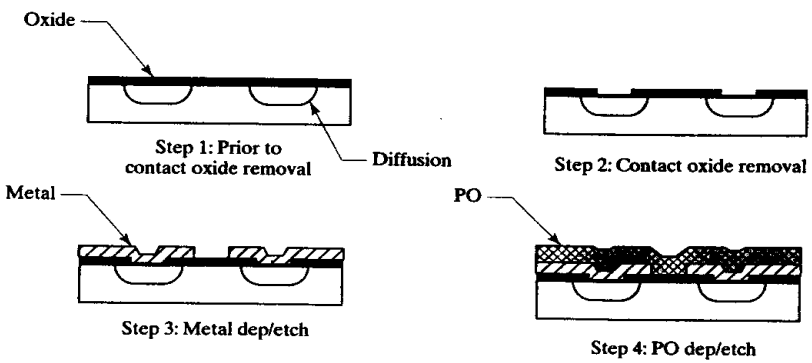


FIGURE 2.24 Formation of a single-level metal system.

Exposed aluminum wiring is vulnerable to mechanical damage and chemical corrosion. An oxide or nitride film deposited over the completed wafer serves as a *protective overcoat* (PO). This layer acts as a conformal seal similar in principle to the plastic conformal coatings sometimes applied to printed circuit boards. Windows etched through the overcoat expose selected areas of the aluminum metallization so that bondwires can be attached to the integrated circuit.

The process illustrated in Figure 2.24 fabricates only a single aluminum layer. Additional layers of metallization can be sequentially deposited and patterned to form a multilevel metal system. Multiple metal layers increase the cost of the integrated circuit, but they allow denser packing of components and therefore reduce the overall die size. The savings in die area often compensate for the cost of the extra processing steps. Multiple metal layers also simplify interconnection and reduce layout time.

CMOS processes frequently employ low-resistivity polysilicon to form the gate electrodes of self-aligned MOS transistors. This material can serve as a free additional layer of interconnect. Even the lowest-sheet poly still has many times the resistance of aluminum, so the designer must take care to avoid routing high-current or high-speed signals in poly. Advanced processes may add a second and even a third layer of polysilicon. These additional layers are used to fabricate different types of MOS transistors, to form the plates of capacitors, and to construct polysilicon resistors. Each of these additional poly layers can be pressed into service as another layer of interconnect.

2.6.1. Deposition and Removal of Aluminum

Most metallization systems employ aluminum or aluminum alloys to form the primary interconnection layers. Aluminum conducts electricity almost as well as copper or silver, and it can be readily deposited in thin films that adhere to all of the materials used in semiconductor fabrication. A brief period of heating will cause the aluminum to alloy into the silicon to form low-resistance contacts.

Aluminum is usually deposited by *evaporation* using an apparatus similar to the one shown in Figure 2.25. The wafers are mounted in a frame that holds their exposed surfaces toward a crucible containing a small amount of aluminum. When the crucible is heated, some of the aluminum evaporates and deposits on the wafer surfaces. A high vacuum must be maintained throughout the evaporation system to prevent oxidation of the aluminum vapor prior to its deposition upon the wafers. The illustrated evaporation system can only handle pure aluminum, but somewhat more complex systems can also evaporate selected aluminum alloys.

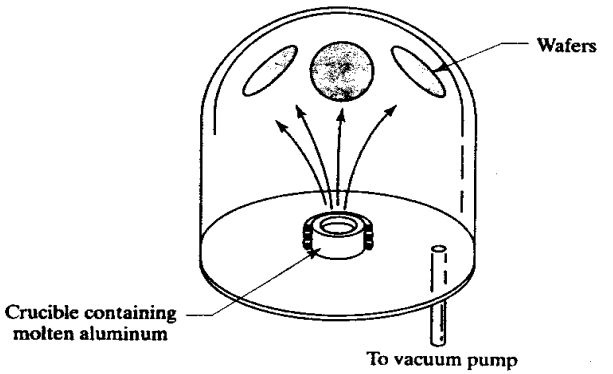


FIGURE 2.25 Simplified diagram of an aluminum evaporation apparatus.

Aluminum and silicon alloy at moderate temperatures. A brief period of heating will form an extremely thin layer of aluminum-doped silicon beneath the contact openings. This process, called *sintering,* can achieve Ohmic contact to P-type silicon because aluminum acts as an acceptor. The aluminum-silicon alloy forms a shallow, heavily doped P-type diffusion that bridges between the metal and the P-type silicon. Less obviously, Ohmic contact also occurs when aluminum touches heavily doped N-type silicon. Junctions form beneath these contacts, but their depletion regions are so thin that carriers can surmount them by quantum tunneling. Rectification will occur if the donor concentration falls too low, so Ohmic contact cannot be established directly between aluminum and lightly doped N-type silicon. The addition of a shallow N+ diffusion will enable Ohmic contact to these regions.

Sintering causes a small amount of aluminum to dissolve in the underlying silicon. Some silicon simultaneously dissolves in the aluminum metal, eroding the silicon surface. Some diffusions are so thin that erosion can punch entirely through them, causing a failure mechanism called *contact spiking.* Historically this was first observed in conjunction with the emitter diffusion of NPN transistors, so it is also called *emitter punchthrough.*[21] Contact spiking can be minimized by replacing pure aluminum metallization with an aluminum-silicon alloy. If the deposited aluminum is already saturated with silicon, then—at least in theory—it cannot dissolve any more. In practice, the silicon content of the alloy tends to separate during sintering to leave an unsaturated aluminum matrix. Careful control of sinter time and temperature will minimize this effect.

Another failure mechanism was encountered in high-density digital logic. As the dimensions of the integrated circuits were progressively reduced, the current density flowing through the metallization increased. Some devices eventually exhibited open-circuit metallization failures after many thousands of hours of operation at elevated temperatures. When the faulty units were examined, some of their leads contained unexpected breaks. These were eventually found to result from a failure mechanism called *electromigration.*[22] Carriers flowing through the metal collide with the lattice atoms. At current densities in excess of several million amps per square centimeter, these impacts become so frequent that the metal atoms begin to move. The displacement of the atoms causes voids to form between individual grains of the polycrystalline metal aggregate. Eventually these voids grow together to form a gap across the entire lead, causing an open-circuit failure (Section 4.1.2). The addition of a fraction of a percent of copper to the aluminum alloy improves electromigration resistance by an order of magnitude. Most modern metal systems therefore employ either aluminum-copper-silicon or aluminum-copper alloys.

2.6.2. Refractory Barrier Metal

The feature sizes of integrated circuits have steadily shrunk as ever-increasing numbers of components have been packed into approximately the same amount of silicon real estate. In order to obtain the necessary packing density, the sidewalls of contact and via openings have become increasingly steep. Evaporated aluminum does not deposit isotropically; the metal thins where it crosses oxide steps (Figure 2.26A). Any reduction in the cross-sectional area of a lead raises the current density and accelerates electromigration. A variety of techniques have been developed to improve step coverage on very steep sidewalls like those formed by reactive ion etching of thick oxide films.

---

[21] M. D. Giles, "Ion Implantation," in S. M. Sze, ed., *VLSI Technology* (New York: McGraw-Hill, 1983), pp. 367–369.

[22] J. R. Black, "Physics of Electromigration," *Proc. 12th Reliability Phys. Symp.*, 1974, p. 142.

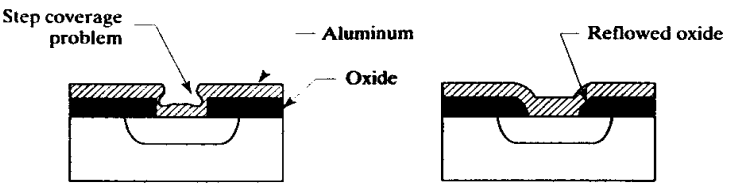


FIGURE 2.26 Step coverage of evaporated aluminum without reflow (A) and with reflow (B).

The step coverage of evaporated aluminum can be greatly increased by moderating the angle of the sidewalls. This can be achieved by heating the wafer until the oxide melts and slumps to form a sloped surface. This process is called *reflow* (Figure 2.26B). Pure oxide melts at too high a temperature to allow reflow, so phosphorus and boron are added to the oxide to reduce its melting point. The resulting doped oxide film is called either a *phosphosilicate glass* (PSG) or a *borophosphosilicate glass* (BPSG), depending on the choice of additives.

Reflow cannot be performed after aluminum has been deposited, because it cannot tolerate the temperatures required to soften PSG or BPSG. Therefore, while reflow can help improve the step coverage of first-level metal, it must be supplemented by other techniques in order to successfully fabricate a multilevel metal system. One option consists of using metals that deposit isotropically upon steeply sloped sidewalls, such as molybdenum, tungsten, or titanium. These *refractory barrier metals* possess extremely high melting points and are thus unsuited for evaporative deposition. A low-temperature process called *sputtering* can successfully deposit them. Figure 2.27 shows a simplified diagram of a sputtering apparatus. The wafers rest on a platform inside a chamber filled with low-pressure argon gas. Facing them is a plate of refractory barrier metal forming one of a pair of high-voltage electrodes. Argon atoms bombard the refractory metal plate. This knocks atoms loose that then deposit on the wafers to form a thin metallic film.

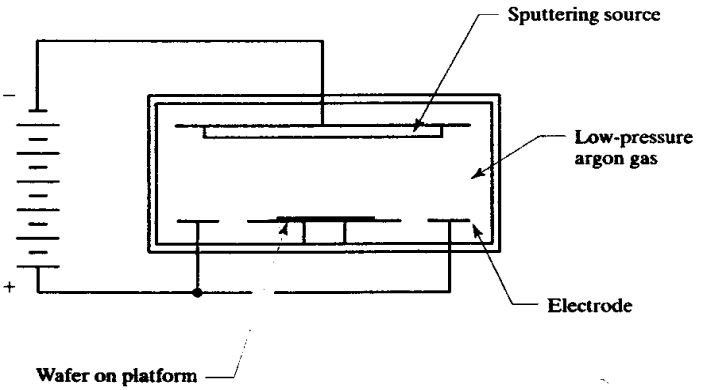


FIGURE 2.27 Simplified diagram of a sputtering apparatus.

The sputtered refractory barrier metal film not only provides superior step coverage, but also virtually eliminates emitter punchthrough.[23] If step coverage were the only criterion for choosing a metal system, then aluminum could be entirely replaced by refractory barrier metal. Unfortunately, refractory metals have relatively

[23] T. Hara, N. Ohtsuka, K. Sakiyama, and S. Saito, "Barrier Effect of W-Ti Interlayers in Al Ohmic Contact Systems," *IEEE Trans. Electron Devices*, Vol. ED-34, #3, 1987, pp. 593–597.

high resistivities and cannot be deposited in thick films as easily as aluminum can. Most metal systems therefore employ a sandwich of both materials. A thin layer of refractory metal beneath the aluminum ensures adequate step coverage in the contacts where the aluminum metal drastically thins. Elsewhere the aluminum reduces the electrical resistance of the metal leads. The relatively short sections of refractory barrier metal in the contacts do not contribute much resistance to the overall interconnection system.

Refractory barrier metals are extremely resistant to electromigration, so the thinning of aluminum on the sidewalls of contacts and vias does not represent an electromigration risk. Refractory barrier metal also tends to suppress classical electromigration failures by bridging any open circuits that develop in the aluminum metallization. Aluminum displaced by electromigration can still short adjacent leads, so refractory barrier metal cannot be relied on to supplement the current-carrying capacity of aluminum wiring except on the sidewalls of contact and via openings.

As mentioned, refractory barrier metal virtually eliminates emitter punchthrough. The degree of alloying between silicon and refractory metals is negligible and the aluminum cannot penetrate the barrier metal to contact the silicon. Most refractory barrier metal systems therefore employ aluminum-copper alloys rather than aluminum-copper-silicon, because aluminum-silicon alloying cannot occur.

2.6.3. Silicidation

Another modification of the standard metallization flow involves the addition of a silicide. Elemental silicon reacts with many metals, including platinum, palladium, titanium, and nickel, to form compounds of definite composition. These *silicides* can form both low-resistance Ohmic contacts and, in the case of certain silicides, stable rectifying Schottky barriers. Thus silicidation not only improves contact resistance—which can be a problem with barrier metal systems—but also allows the formation of Schottky diodes at no extra cost. Silicides have much lower resistivities than even the most heavily doped silicon, so they can also be used to reduce the resistance of selected silicon regions. Many MOS processes employ silicided poly (also called *clad poly*) to form the gates of high-speed MOS transistors. Some of these processes also clad the source/drain regions of the transistors to reduce their resistance. Since most silicides are relatively refractory, their deposition does not preclude subsequent high-temperature processing. Silicided gates can thus be used to form self-aligned source/drain regions.

Figure 2.28 shows the steps required to deposit a platinum silicide layer in selected regions of the wafer. Immediately after the contacts are opened, a thin film of

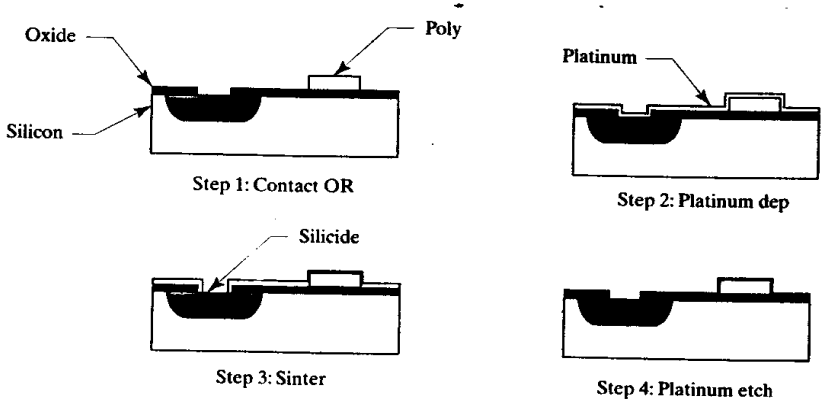


FIGURE 2.28 Silicidation process, showing both silicided contacts and silicided poly.

platinum metal is deposited across the entire wafer. The wafer is then heated to cause the portions of the platinum film in contact with silicon to react to form platinum silicide. The unreacted platinum can be removed using a mixture of acids called aqua regia. This procedure silicides both contact openings and any exposed polysilicon. If desired, an additional masking step can select which regions should receive silicide. Processes employing clad poly must incorporate a silicide block mask to fabricate polysilicon resistors. If this were not done, silicidation would turn all of the poly into a low-resistance material.

A typical silicided metal system consists of a lowermost layer of platinum silicide, an intermediate layer of refractory barrier metal,[24] and a topmost layer of copper-doped aluminum. The resulting sandwich exhibits low electrical resistance, high electromigration immunity, stable contact resistance, and precisely controlled alloying depths. The three layers required to obtain all of these benefits are more costly than a simple aluminum alloy metallization, but the performance benefits are substantial.

2.6.4. Interlevel Oxide, Interlevel Nitride, and Protective Overcoat

Figure 2.29 shows a cross section of a typical modern metallization system. The first layer of material above the silicon consists of thermally grown oxide. Upon this oxide lies a patterned polysilicon layer that will eventually form the gates of MOS transistors. On top of this poly lies a thin deposited oxide layer called a *multilevel oxide* (MLO) that serves to insulate the poly and to thicken the thermal oxide layer. Contact openings are etched through the MLO and thermal oxide to contact the silicon, and through the MLO to contact the poly. Following reflow, the contact openings are silicided to reduce contact resistance. Above the MLO lies the first layer of metal, consisting of a thin film of refractory barrier metal and a much thicker layer of copper-doped aluminum. Above the first metal layer lies another deposited oxide layer called an *interlevel oxide* (ILO), which insulates the first metal from the overlying second metal. Vias are etched through the ILO. On top of this lies the second layer of metal, again consisting of refractory barrier metal and copper-doped aluminum. The topmost and final layer consists of a compressive nitride film, which serves as a *protective overcoat* (PO). This metallization system has a total of six layers (one poly, two metals, MLO, ILO, and PO) and requires five masking steps (poly, contact, metal-1, via, metal-2, and PO). Some advanced processes may employ as many as three layers of polysilicon and five layers of metal.

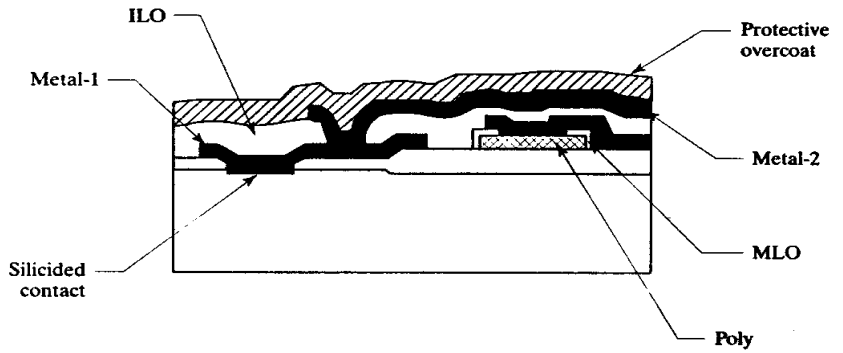


FIGURE 2.29 Cross section of a double-metal, single-poly metallization system.

[24] The addition of a refractory barrier metal prevents the platinum silicide from reacting with the aluminum. This is not required for most refractory silicides; see Sze, pg. 409.

Interlevel oxide layers are normally produced by low-temperature deposition—for example, by the reaction of silane and nitrous oxide or by the decomposition of tetraethoxysilane (TEOS). A relatively thick ILO helps minimize parasitic capacitances between layers of the conductor sandwich, but it can cause step coverage problems in via openings. As previously discussed, reflow is not possible once aluminum has been deposited, so a refractory barrier metal is often used to improve the step coverage of the second metal layer.

An excellent capacitor can be formed between two layers of metal or polysilicon. A thin insulating dielectric deposited between the plates completes the capacitor. The thinner this dielectric, the greater the resulting capacitance per unit area. One technique for forming a capacitor consists of depositing one polysilicon layer, oxidizing this to form a thin dielectric, and depositing a second polysilicon layer to complete the capacitor. Any region where the two poly layers overlap will form a capacitor consisting of two poly plates separated by the thin oxide dielectric. Oxide forms an ideal capacitor dielectric because it is a nearly perfect insulator, and very thin oxide films can be grown with little risk of pinholes or other defects. The capacitance achievable with oxide dielectrics is limited by the rupture voltage of the oxide; the thicker oxide layers required to withstand higher voltages have proportionately smaller capacitances per unit area.

One way to boost the capacitance per unit area for a given operating voltage consists of using a material with a higher dielectric constant. Silicon nitride, with a dielectric constant that is 2.3 times that of oxide, is a common choice for fabricating high capacitance-per-unit-area films. Unfortunately, nitride films are more prone to pinhole formation than are oxide films of equivalent thickness. Therefore oxide and nitride films are sometimes combined to form a stacked dielectric with a dielectric constant between that of oxide and nitride. A typical oxide-nitride-oxide stacked dielectric can achieve a dielectric constant about twice that of oxide.

The protective overcoat consists of a thick deposited oxide or nitride film coating the entire integrated circuit. It insulates the uppermost metal layer from the outside world, so that (for example) a particle of conductive dust will not short two adjacent leads. The overcoat also helps to ruggedize the integrated circuit, a necessary precaution since the aluminum metallization is soft and deforms under pressure. The protective overcoat also helps to block the entrance of contaminants. Both the aluminum metallization and the underlying silicon are vulnerable to certain types of contaminants that can penetrate the plastic encapsulation. A properly formulated protective overcoat can form a barrier to these contaminants. Heavily doped phosphosilicate glasses are sometimes used as protective overcoats, but most modern processes have switched to compressive nitride films, which offer superior mechanical hardness and chemical resistance.

2.7 ASSEMBLY

Wafer fabrication ends with the deposition of a protective overcoat, but there remain a number of manufacturing steps required to complete the integrated circuits. Since most of these steps require less-stringent cleanliness than wafer fabrication, they are usually performed in a separate facility called an *assembly/test site.*

Figure 2.30 shows a diagram of a typical finished wafer. Each of the small squares on the wafer represents a complete integrated circuit. This wafer contains approximately 300 integrated circuit dice arrayed in a rectangular pattern by the step-and-repeat process that created the stepped working plates. A few locations in the array are occupied by process control structures and test dice rather than actual integrated circuits.

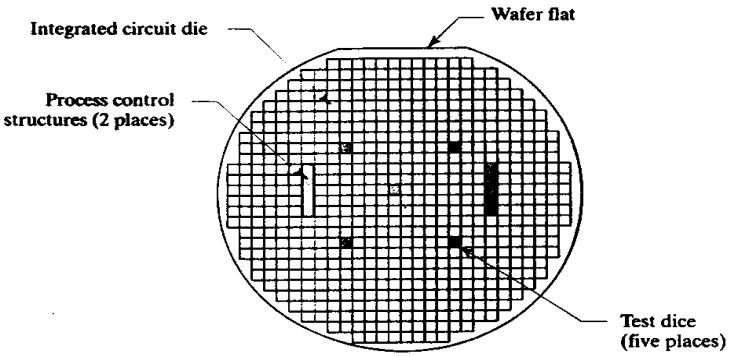


FIGURE 2.30 Pattern for a typical wafer created from a stepped working plate.

Process control structures consist of extensive arrays of transistors, resistors, capacitors, and diodes, as well as more specialized structures such as strings of contacts and vias. The wafer fab uses these structures to evaluate the success or failure of the manufacturing process. Automated testing equipment gathers data on every wafer, and any that fail to meet specifications are discarded. The data are also analyzed for statistical trends so that corrections can be implemented before the variances become large enough to cause yield losses. The process control structures are standardized, and the same ones are used for a wide range of products.

Test dice are used by design engineers to evaluate prototypes of an integrated circuit. Unlike process control structures, test dice are specific to a given product and in most cases are actually variations on the layout of the integrated circuit. A dedicated test metal mask allows probing of specific components and subcircuits that would be difficult to access on the finished die. Sometimes a test contact or protective overcoat mask is also used, but in almost every case the test die shares the same diffusion masks as the integrated circuit. Test dice are normally created by adding a few more layers (e.g., test metal, test nitride) to the database containing the layout of the integrated circuit. These layers create a separate set of reticles that are used to expose a few selected spots on the stepped working plate. The wafer in Figure 2.30 contains only five test die locations. These locations become unnecessary when testing has been completed. Sometimes a new set of masks is created that replaces the test dice with product dice to gain an extra percent or two of yield. In other cases, the tiny increase in die yield cannot justify fabrication of new masks, so the test dice remain on production material.

Figure 2.30 depicts a wafer produced from a set of stepped working plates. Wafers created by *direct-step-on-wafer* (DSW)[25] processing rarely include any test dice because at least one test die must be included in every exposure. This would result in twenty or more test dice per wafer, which would consume a corresponding amount of area. If test dice are included in a DSW design, then the production mask set will almost certainly be modified to replace them with product dice to improve the die yield.

As mentioned previously, all completed wafers are tested to determine whether the processing was performed correctly. If the wafers pass this test, then each die is

[25] The acronym *DSW* has also been used to stand for *direct slice write*, a process by which a scanned electron beam directly exposes the photoresist on a wafer. This process, more commonly called *direct-write-on-wafer* (DWW), is strictly of academic interest because it is too slow to have any practical application in silicon processing. However, it is frequently used to fashion photomasks.

individually tested to determine its functionality. The high-speed automated test equipment typically requires less than three seconds to test each die. The percentage of good dice depends on many factors, most notably the size of the die and the complexity of the process used to create it. Most products yield better than 80% and some yield in excess of 90%. High yields are obviously desirable because every discarded die represents lost profit. The equipment that tests the wafers also marks those that fail the test. Marking is usually done by placing a drop of ink on each defective die, but some modern systems eliminate the need for inking by remembering the location of the bad dice electronically.

Wafer-level testing, or *wafer probing,* requires contact to specific locations on the interconnection pattern of the integrated circuit. These locations are exposed through holes in the protective overcoat, allowing contact to be made with the help of an array of sharp metal needles, or *probes.* These probes are mounted on a board called a *probe card.* The automated test machine lowers the probe card until electrical continuity is established. The integrated circuit is tested, the card is lifted, and positioning servos move the wafer to align the next die underneath the probes.

Once the wafer has been completely tested, the individual dice are sawn apart using a diamond-tipped sawblade. Another automated system then selects the good dice from the scribed wafer for mounting and bonding. The rejected dice (including the remains of the process control structures and the test dice) are discarded.

2.7.1. Mount and Bond

Many manufacturers now offer unmounted integrated circuit dice, but the sales of such *bare dice* are seldom large. Most customers simply do not have the equipment or expertise needed to handle bare dice, let alone to package them. Packaging therefore falls in the province of integrated circuit manufacturing.

The first step in packaging an integrated circuit is mounting it on a *leadframe.* Figure 2.31 shows a somewhat simplified diagram of a leadframe for an eight-pin *dual-in-line package* (DIP), complete with a chip mounted on it. The leadframe itself consists of a rectangular *mount pad* that holds the die and a series of lead fingers that will eventually be trimmed to form the eight leads of the DIP. Leadframes usually come in strips, so several dice can be handled as a single assembly. They are either stamped out of thin sheets of metal, or they are etched using photographic techniques similar to those used to pattern printed circuit boards. The lead frame usually consists of copper or a copper alloy, often plated with tin or a tin-lead alloy. Copper is not an ideal material for leadframes because it has a different coefficient

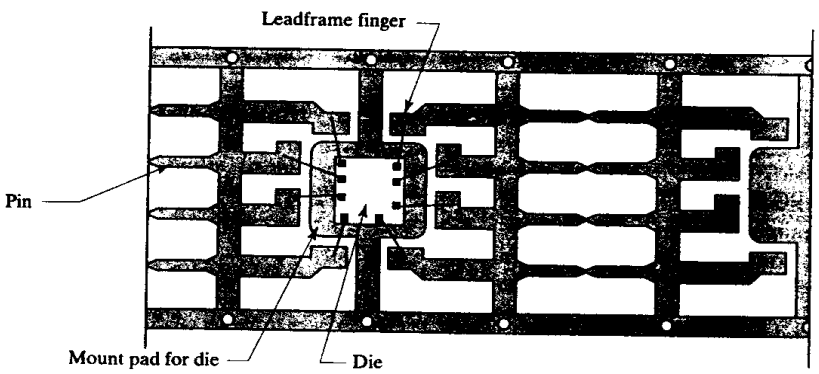


FIGURE 2.31 Simplified diagram of a leadframe for an 8-pin DIP.

of thermal expansion than silicon. As the packaged part is heated and cooled, differential expansion of the die and the leadframe sets up mechanical stresses injurious to the performance of the die. Unfortunately, most of the materials that possess coefficients of expansion similar to silicon have inferior mechanical and electrical properties. Some of these materials are occasionally used for low-stress packaging of specialty parts; a nickel-iron alloy called *Alloy-42* is probably the most commonly encountered (Section 7.2.6).

The die is usually mounted to the leadframe using an epoxy resin. In some cases, the resin may be filled with silver powder to improve thermal conductivity. Epoxy is not entirely rigid, and this helps reduce the stresses produced by thermal expansion of the leadframe and die. Alternate methods exist that provide superior thermal union between the silicon and the leadframe, but at the cost of greater mechanical stress. For example, the backside of the die can be plated with a metal or metal alloy and soldered to the leadframe. Alternatively, a rectangle of gold foil called a *gold preform* can be attached to the leadframe; heating the die causes it to alloy with the gold preform to create a solid mechanical joint. Solder connections and gold preforms both allow excellent thermal contact between the die and the leadframe. Both also produce an electrical connection that can be used to connect the substrate of the die to a pin. Conductive epoxies improve thermal conductivity, but they cannot always be trusted to provide electrical connectivity.[26]

After the dice are mounted on leadframes, the next step is attaching bondwires to them. Bonding can only be performed in areas of the die where the metallization is exposed through openings in the protective overcoat; these locations are called *bondpads*. The probe card used for wafer probing makes contact to the bondpads for purposes of testing, but the probes may also make contact to a few pads that will not receive bondwires. Those pads reserved for testing purposes are usually called *testpads* to distinguish them from actual bondpads. Testpads are often made smaller than bondpads since probe needles can usually land in a smaller zone than can bondwires.

Bonding is performed by high-speed automated machines that use optical recognition to determine the locations of the bondpads. These machines typically employ one-mil (25μm) gold wire for bonding, although gold wires as small as 0.8 mil or as large as 2.0 mil are in common use. Aluminum wires up to ten mils in diameter can also be employed, although these require rather different bonding machinery. Only one diameter and type of wire can be bonded at a time, so few dice use more. The most common arrangement consists of one-mil gold bondwires on all pads. Multiple one-mil wires bonded in parallel can carry higher currents or provide lower resistances without requiring a second bonding pass for larger-diameter wire.

The most common technique for bonding gold wire is called *ball bonding*.[27] Since aluminum wire cannot be ball-bonded, an alternate technique called *wedge bonding* has been developed for it. Figure 2.32 shows the essential steps of the ball-bonding process.

The bonding machine feeds the gold wire through a slender tube called a *capillary*. A hydrogen flame melts the end of the wire to form a small gold sphere, or *ball* (Figure 2.32, Step 1). Once a ball has been formed, the capillary presses down against the bondpad. The gold ball deforms under pressure, and the gold and aluminum fuse together to form a weld (Step 2). Next, the capillary lifts and moves to the vicinity of

---

[26] R. L. Opila and J. D. Sinclair, "Electrical Reliability of Silver Filled Epoxies for Die Attach." *23rd International Reliability Physics Symp.*, 1985, pp. 164–172.

[27] B. G. Streetman, *Solid State Electronic Devices*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1980), pp. 368–370.

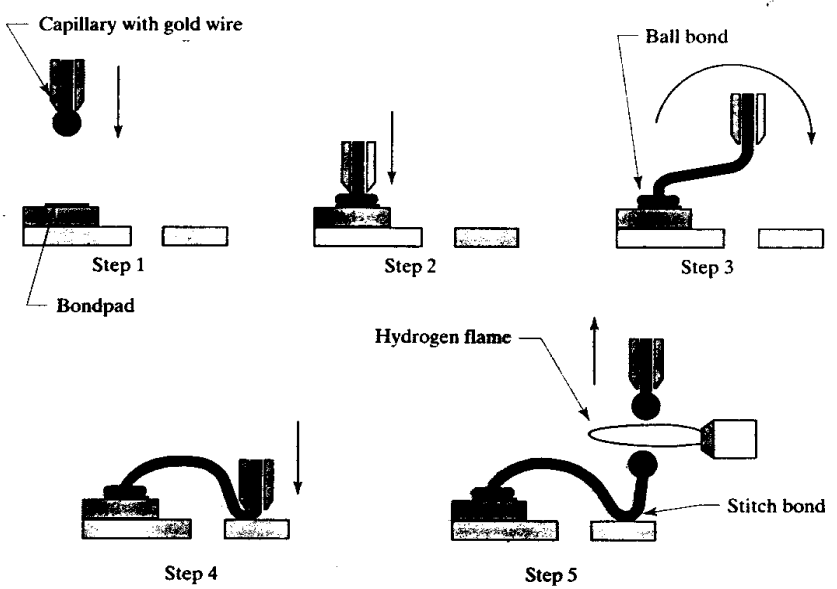


FIGURE 2.32 Steps in the ball-bonding process.

the lead finger (Step 3). The capillary again descends, smashing the gold wire against the lead finger. This causes the gold to alloy to the underlying metal to produce a weld (Step 4). Since no ball is present at this point, the resulting bond is called a *stitch bond* rather than a ball bond. Finally, the capillary lifts up from the lead finger and the hydrogen flame passes through the wire, causing it to fuse into two (Step 5). The bond is now complete and another ball has been formed on the wire protruding from the capillary, allowing the process to be repeated. An automated bonding machine can perform these steps ten times a second with great precision. The extreme speed and unerring accuracy of these machines produce huge economies of scale, and the entire bonding process costs no more than a penny or two.

Aluminum wire cannot be ball-bonded because the hydrogen flame would ignite the fine aluminum wire. Instead, a small, wedge-shaped tool is used to supplement the capillary. When the capillary brings the wire into proximity with the bondpad, the tool smashes it against the pad to create a stitch bond. The process is repeated at the lead finger, and the tool is then held down against the lead finger while the capillary moves up. The tension created in the aluminum wire snaps it at its weakest point, immediately adjacent to the weld. The process can then be repeated as many times as necessary.

The ball-bonding process requires a square bondpad approximately three times as wide as the diameter of the wire. Thus a one-mil gold wire can be attached to a square bondpad about three mils across. Wedge bonding is more selective, and usually requires bondpads that are rectangular in shape. These bondpads must lie in the same direction as the wedge tool. They are typically twice as wide and four times as long as the wire is thick. The exact rules for wedge bondpads can become quite complex, particularly for thicker aluminum wires.

Figure 2.31 shows a die mounted on a leadframe after the bonding process is complete. The bondwires connect various bondpads to their respective leads. Although the wires are quite small compared to the pins, each is still capable of carrying an amp of current.

2.7.2 Packaging

The next step in the assembly process is injection molding. A mold is clamped around the leadframe and heated plastic resin is forced into the mold from below. The plastic wells up around the die, lifting the wires away from it in gentle loops. Injection from the side or from the top usually smashes the wires against the integrated circuit and is therefore impractical. The plastic resin employed for integrated circuits cures rapidly at the temperatures used in molding and, once cured, it forms a rigid block of plastic.

When the molding process is complete, the leads are trimmed and formed to their final shapes. This is done in a mechanical press using a pair of specially shaped dies that simultaneously trim away the links between the individual leads and bend them to the required shape. Depending on the material of the leadframe, solder dipping or plating may be required to prevent oxidation and contamination of the pin surfaces. The completed integrated circuits are now labeled with part numbers and other designation codes (these usually include a code identifying the date of manufacture and the lot number). The completed integrated circuits are tested again to ensure that they have not been damaged by the packaging process. Finally, the completed devices are packaged in tubes, trays, or reels for distribution to customers.

2.8 SUMMARY

Modern semiconductor processing takes advantage of the properties of silicon to manufacture inexpensive integrated circuits in huge volumes. Photolithography allows the reproduction of intricate patterns hundreds or thousands of times across each wafer, leading to enormous economies of scale.

Junctions can be formed by one of three means: epitaxial deposition, diffusion, or ion implantation. Low-pressure chemical-vapor-deposited (LPCVD) epitaxial layers can produce extremely high-quality silicon films with precisely controlled dopant concentrations. Diffusion of dopants from a surface source allows the formation of vast numbers of junctions using only a single photolithographic masking step. Ion implantation allows similar but more costly patterning of junctions with superior control of doping levels and distributions.

Many materials can also be deposited on the surface of the wafer. These include polycrystalline silicon (poly), oxide, nitride, and any of numerous metals and metal alloys. Typical semiconductor processes combine several diffusions into the bulk silicon with several depositions of materials onto the resulting wafer. The next chapter examines how the various techniques of semiconductor fabrication are combined to manufacture three of the most successful integrated circuit processes.

2.9 EXERCISES

2.1. When pressure is applied to the center of an unknown wafer, it splits into six segments. What can be definitely concluded from this observation? What may be reasonably conjectured?

2.2. Draw a diagram similar to that in Figure 2.4 illustrating the relationship between the (100) and (110) planes of a cubic crystal (refer to Appendix B, if necessary).

2.3. Suppose a photomask consists of a single opaque rectangle on a clear background. A negative resist is used in combination with this mask to expose a sensitized wafer. Describe the pattern of photoresist left on the wafer after development.

2.4. Suppose a wafer is subjected to the following processing steps:

a. Uniform oxidation of the entire wafer surface.

b. Opening of an oxide window to the silicon surface.

c. An additional period of oxidation.
d. Opening of a smaller oxide window in the middle of the region patterned in step b.
e. An additional period of oxidation.

Draw a cross section of the resulting structure, showing the topography of both the silicon and the oxide surfaces. The drawing need not be made to scale.

2.5. Suppose a wafer is uniformly doped with equal concentrations of boron and phosphorus atoms. After a prolonged oxidation, will the surface of the silicon be N-type or P-type? Why?

2.6. Suppose that a wafer is uniformly doped with 10^{16} atoms/cm^3 of boron. This wafer is then subjected to the following processing steps:
a. Uniform oxidation of the entire wafer surface.
b. Opening of an oxide window to the silicon surface.
c. Deposition of boron and phosphorus, each at a source concentration of 10^{20} atoms/cm^3, using a fifteen-minute deposition and a one-hour drive at 1000°C.

Assuming that the two dopants do not interact with each other or with the oxide, draw a cross section of the resulting structure. Indicate the approximate depths of any junctions formed.

2.7. Phosphorus is diffused into a lightly doped wafer through an oxide window 5μm square. If the resulting junction is 2μm deep, then what is the width of the phosphorus diffusion at the surface?

2.8. Most ion implantation systems position the accelerator so that the ion beam impacts the wafer surface at a slight angle (often 7°). Explain the reason for this feature.

2.9. If the surface of the oxide layer covering a wafer is ground perfectly smooth, different regions of the wafer still exhibit different colors, but the NBL shadow vanishes. Explain these observations.

2.10. Draw a cross section of the following metallization system:
a. 1μm-wide contacts through 5kÅ oxide silicided with 2kÅ platinum silicide.
b. First-level metal consisting of 2kÅ RBM and 6kÅ copper-doped aluminum.
c. 1μm-wide vias 3kÅ deep through highly planarized ILO.
d. Second-level metal consisting of 2kÅ RBM and 10kÅ copper-doped aluminum.
e. 10kÅ protective overcoat.

Assume that the silicide surface is level with the surrounding silicon surface, and that the aluminum metal thins 50% on the sidewalls of the contacts and vias. The drawing should be made to scale.

2.11. Suppose that a die measures 60 by 80 mils, where one mil is one thousandth of an inch. Approximately how many of these dice could be fabricated on a 150mm-diameter wafer? Assuming that 70% of these potential dice actually work, and that a finished wafer costs $250, compute the cost of each functional die.

3 Representative Processes

Semiconductor processing has evolved rapidly over the past fifty years. The earliest processes produced only discrete components, mainly switching diodes and bipolar transistors. The first practical integrated circuits appeared in 1960.[1] These consisted of a few dozen bipolar transistors and diffused resistors connected to form simple logic gates. By modern standards, these early integrated circuits were terribly slow and inefficient. Refinements were soon made, and by the mid-1960s bipolar integrated logic offered clear advantages over discrete logic. The first analog integrated circuits appeared at about the same time; these consisted of matched transistor arrays, operational amplifiers, and voltage references. The standard bipolar process that was created to support these products remains in use today.

Integrated bipolar logic was fast but power-hungry. MOS integrated circuits held out the promise of a low-power alternative, but the metal-gate MOS processes of the 1960s suffered from unpredictable threshold voltage shifts. This problem was eventually conquered through the development of polysilicon-gate MOS processes in the early 1970s. MOS logic soon replaced bipolar logic and created vast new markets for microprocessors and dynamic RAM chips. Analog CMOS circuits of this era touted greatly reduced operating currents but provided only mediocre performance, so standard bipolar remained the process of choice for high-performance analog integrated circuits.

By the mid-1980s, customers were demanding the integration of both digital and analog functions onto a single mixed-signal integrated circuit. A new generation of merged bipolar-CMOS (BiCMOS) processes were soon developed specifically for mixed-signal design. Although these processes are complex and costly, they offer a level of performance unachievable by other means. The world of analog integrated circuits is dominated by these three processes: standard bipolar, polysilicon-gate CMOS, and analog BiCMOS. This chapter will analyze the implementation of a representative process of each type.

[1] J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. on Electron Devices*, Vol. ED-23, #7, 1976, pp. 648–654.

3.1 STANDARD BIPOLAR

Standard bipolar was the first analog integrated circuit process. Over the years, it has produced many classic devices, including the 741 operational amplifier, the 555 timer, and the 431 voltage reference. Even though these parts represent thirty-year-old technology, they are still manufactured in vast quantities today.

Standard bipolar is seldom used for new designs. CMOS offers lower supply currents, BiCMOS provides superior analog performance, and various advanced bipolar processes yield faster switching speeds. But the knowledge gained through first developing and then refining standard bipolar will never become obsolete. The same devices reappear in every new process, along with many of the same parasitic mechanisms, design tradeoffs, and layout principles. This chapter will therefore begin with an overview of standard bipolar.

3.1.1. Essential Features

Standard bipolar was shaped by a conscious decision to optimize the NPN transistor at the expense of the PNP. This decision rested on the observation that the NPN transistor employs electron conduction while the PNP transistor relies on hole conduction. The lower mobility of holes reduces both the beta and the switching speed of PNP transistors. Given equivalent geometries and doping profiles, an NPN will outperform a PNP by more than 2:1. Several additional processing steps are required to optimize both types of transistors simultaneously, so early processes optimized NPN transistors and avoided PNP transistors altogether. This decision met the requirements of bipolar logic, consisting as it does of NPN transistors, resistors, and diodes. When analog circuits were first constructed using standard bipolar, several types of PNP transistors were cobbled together from existing process steps. Although these transistors performed relatively poorly, they sufficed to design many useful circuits.

The standard bipolar process employs *junction isolation* (JI) to prevent unwanted currents from flowing between devices that are formed on the same substrate.[2] The components reside in a lightly doped N-type epitaxial layer deposited on top of a lightly doped P-type substrate (Figure 3.1). A deep-P+ *isolation diffusion* driven down to contact the underlying substrate provides isolation between components. Regions of N-epi separated from one another by isolation are called *tanks* or *tubs*. If the isolation is biased to a potential equal to or below the lowest-voltage tank, then reverse-biased junctions surround every tank. The substrate forms the floor of these tanks, and isolation diffusions form their sidewalls.

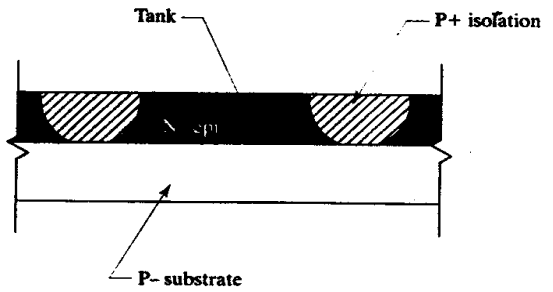


FIGURE 3.1 Cross section of the junction isolation system employed for standard bipolar.

[2] R. N. Noyce, U.S. Patent #2,981,877, 1961.

Junction isolation has several significant drawbacks. The reverse-biased isolation junctions exhibit enough capacitance to slow the operation of many circuits. High temperatures can cause significant leakage currents, as can exposure to light or ionizing radiation. Unusual operating conditions can also forward bias the isolation junctions and inject minority carriers into the substrate. Despite these difficulties, junction-isolated processes can successfully fabricate most circuits. Junction isolation is also considerably cheaper than any of its alternatives.

3.1.2. Fabrication Sequence

The baseline standard bipolar fabrication sequence consists of eight masking operations. The significance of each step can be illustrated best by presenting the entire flow from starting material to finished wafer. Representative cross sections will be used to illustrate each step. When examining these cross sections (and all others in this text), keep in mind that the vertical scale of the drawings has been exaggerated by a factor of two to five for clarity. The lateral dimensions of a typical integrated device are so much greater than its vertical dimensions that a true-scale diagram would be virtually illegible. The cross sections therefore exaggerate the vertical scale by a factor of two to five. The substrate is also much thicker than depicted; the additional silicon serves to strengthen the wafer against warping and breakage.

Starting Material

Standard bipolar integrated circuits are fabricated on a lightly doped (111)-oriented P-type substrate. The wafers are usually cut several degrees off-axis to minimize distortion of the NBL shadow (*pattern distortion*).[3] The use of (111) silicon helps suppress a parasitic PMOS transistor that is inherent in the standard bipolar process. The N-epi forms the backgate of this parasitic, while a lead crossing the field oxide above the tank acts as its gate electrode. A base region within the tank forms the source, and the drain consists of the P+ isolation (Figure 3.2). When the base diffusion is biased to a high voltage relative to the metal lead, a channel forms and allows current to flow from the base to the isolation. The threshold voltage of an MOS transistor formed under thick field oxide is called a *thick-field threshold*. The use of (111) silicon artificially elevates the PMOS thick-field threshold by introducing positive surface-state charges along the oxide-silicon interface.

N-Buried Layer

The first processing step consists of growing a thin layer of oxide across the wafer. Photoresist spun onto this oxide is patterned using the N-buried layer (NBL) mask.

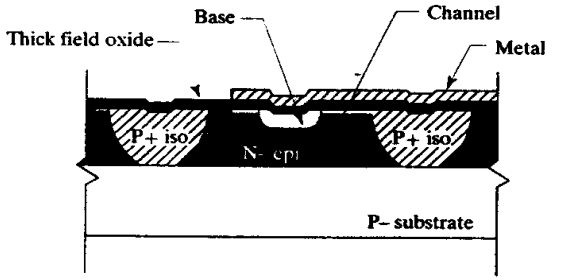


FIGURE 3.2 Parasitic PMOS formation in standard bipolar.

[3] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology* (Reading, MA: Addison-Wesley, 1994), p. 331.

After an oxide etch opens windows to the silicon surface, ion implantation or thermal deposition transfers an N-type dopant into the wafer. The N-buried layer customarily consists of either arsenic or antimony because the low diffusivities of these elements limit up-diffusion during subsequent processing. The brief drive following the deposition serves two purposes: first, it anneals out lattice damage; and second, it grows a small amount of oxide that forms a slight discontinuity at the silicon surface (Figure 3.3). This discontinuity will later produce an NBL shadow to which other masks can align.

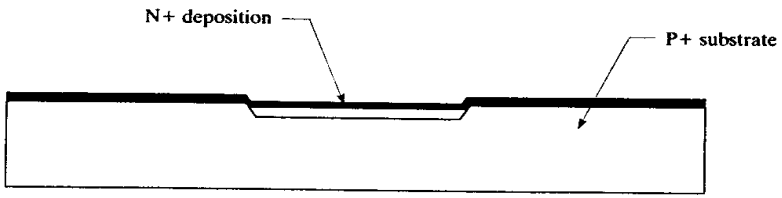


FIGURE 3.3 Wafer after anneal of NBL implant.

Epitaxial Growth

The oxide layer that remains on the wafer is stripped prior to the growth of some 10 to 25μm of lightly doped N-type epi. Surface discontinuities propagate upward during epitaxial deposition at approximately a 45° angle along the axis. Upon completion of epitaxial growth, the NBL shadow will have shifted laterally a distance approximately equal to the thickness of the epi (Figure 3.4).

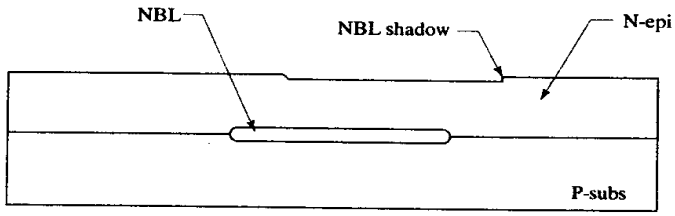


FIGURE 3.4 Wafer after epitaxial deposition. Note the pattern shift exhibited by the NBL shadow.

Isolation Diffusion

The wafer is again oxidized, coated with photoresist, and patterned using the *isolation* mask. This mask must be aligned to the NBL shadow using a deliberate offset to correct for pattern shift. A heavy boron deposition followed by a high-temperature drive forces the isolation diffusion partway down through the epi layer. Oxidation also occurs during this drive, covering the isolation windows with a thin layer of thermal oxide. The drive stops before the isolation junction reaches the substrate, since later high-temperature processing steps (primarily the deep-N+ drive) will force the diffusion the rest of the way down. Figure 3.5 shows the wafer after this partial drive.

Deep-N+

A deep-N+ diffusion (sometimes called a *sinker*) allows low-resistance connection to the NBL. First a photoresist is applied and patterned using the deep-N+ mask. A heavy phosphorus deposition followed by a high-temperature drive forms the deep-N+ sinkers. The drive not only causes the deep-N+ to diffuse down to meet the upward-diffusing NBL but also completes the isolation drive. Sufficient time is

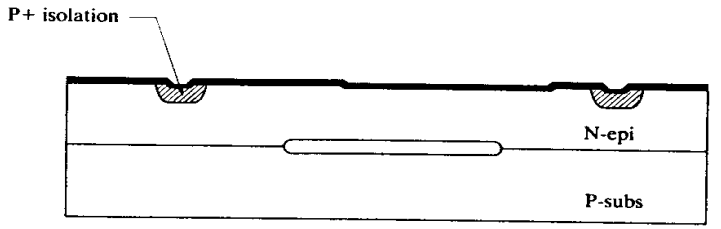


FIGURE 3.5 Wafer after isolation deposition and partial drive.

allowed to overdrive the junctions by about 25%. Without this overdrive, the bottom of the isolation and deep-N+ diffusions would be very lightly doped. The overdrive simultaneously reduces the vertical resistance through both the isolation and the deep-N+ sinkers. The deep-N+ drive also forms the thick field oxide.

Both deep-N+ and isolation diffusions approach their final junction depths during the deep-N+ drive (Figure 3.6). These junctions will diffuse slightly deeper during subsequent processing, but all of the later diffusions are fairly shallow compared to deep-N+ and isolation, and therefore the tank regions appear fully formed in Figure 3.6. The NBL regions are normally spaced some distance inside the isolation diffusion to increase the tank breakdown voltage. Otherwise the N+/P+ junction formed by the intersection of NBL and isolation would avalanche at about 30V.

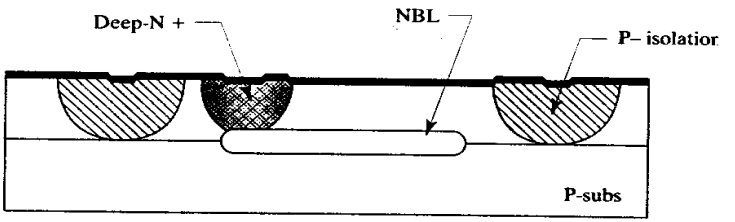


FIGURE 3.6 Wafer after isolation drive.

Base Implant

Next, photoresist spun onto the wafer is patterned using the base mask. An oxide etch clears windows through the field oxide to the silicon surface. A light boron implant conducted through these openings counterdopes the N-epi to form the base regions of the NPN transistors. Ion implantation allows precise control of base doping and thus minimizes process-derived beta variation. The subsequent drive anneals implant damage and sets the base junction depth. Oxide grown during this drive serves as a mask for the subsequent emitter deposition. Base is also implanted across the isolation regions to increase surface doping. This practice, called *base-over-isolation* (BOI), substantially increases the NMOS thick-field threshold without requiring the use of a separate channel stop. Figure 3.7 shows a cross section of the wafer following the base drive.

Emitter Diffusion

The wafer is again coated with photoresist and patterned using the emitter mask. A subsequent oxide etch exposes the silicon surface in regions where NPN emitters will form and in regions where Ohmic contact must be made to the N-epi or the deep-N+ diffusion. A very concentrated phosphorus deposition forms the emitter.

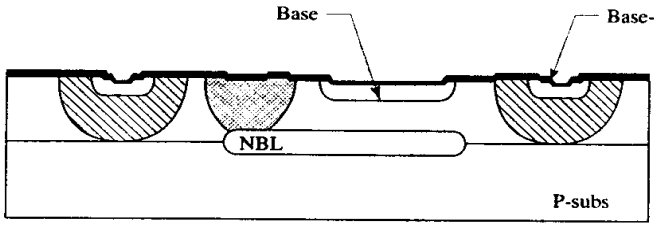


FIGURE 3.7 Wafer after base drive.

A $POCl_3$ source is often used since precise control of emitter doping is unnecessary. A brief drive sets the final emitter junction depth and thereby determines the width of the active base region of the NPN transistors.

An oxide film grown over the emitter diffusion insulates it from subsequent metallization. Some processes employ dry oxidation for this step, but the short oxidation time results in a *thin emitter oxide* vulnerable to electrostatic discharges (Section 4.1.1). Alternatively, a wet oxidation can grow a *thick emitter oxide* that possesses a higher rupture voltage. Figure 3.8 shows a cross section of the wafer after the emitter drive.

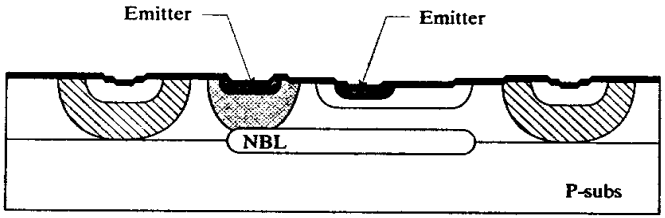


FIGURE 3.8 Wafer after emitter drive.

Many older processes also incorporate an *emitter pilot* step to provide a means of adjusting NPN beta. A dummy wafer that is inserted into the wafer lot before base implant and removed after emitter deposition is used to conduct an experimental emitter drive. By monitoring the performance of the base-emitter junction formed on the pilot wafer, the actual drive can be adjusted on a lot-by-lot basis to target the desired NPN beta.

Contact

All diffusions are now complete. The remaining steps form the metallization and apply the protective overcoat. The first step in this sequence forms contacts to selected diffusions. The wafer is again coated with photoresist, patterned using the contact mask, and etched to expose bare silicon. This process is sometimes called *contact OR*, in which OR stands for *oxide removal*.

Metallization

A layer of aluminum-copper-silicon alloy is evaporated or sputtered across the entire wafer. This metal system typically incorporates 2% silicon to suppress emitter punchthrough and 0.5% copper to improve electromigration resistance. Standard bipolar employs relatively thick metallization, typically at least 10kÅ (1.0μm) thick, to reduce interconnection resistance and decrease vulnerability to electromigration. The metallized wafer is patterned using the metal mask and etched to form the interconnection system.

Protective Overcoat

Next, a thick layer of *protective overcoat* (PO) is deposited across the entire wafer. Compressive nitride protective overcoats provide excellent mechanical and chemical protection. Some processes use a *phosphosilicate-doped glass* (PSG) layer either beneath a compressive nitride or as a replacement for it. Since the deposition of the protective overcoat occurs at moderate temperatures, it also sinters the aluminum metallization.

Finally, a layer of photoresist is applied and patterned using the PO mask. A special etch opens windows through the protective overcoat to expose areas of metallization for bonding. This composes the final fabrication step; the wafer is now complete. Figure 3.9 shows a fully processed wafer (the illustrated cross section does not include a bondpad opening).

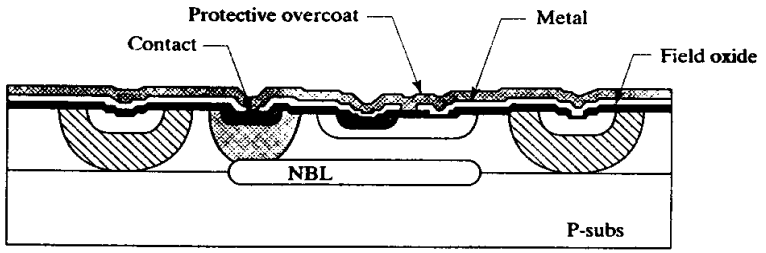


FIGURE 3.9 Completed standard bipolar wafer.

3.1.3. Available Devices

Standard bipolar was originally developed to provide bipolar NPN transistors and diffused resistors. A number of other devices can also be fabricated using the same process steps, including two types of PNP transistors, several types of resistors, and a capacitor.[4] These devices form a basic component set suitable for fabricating a wide variety of analog circuits. Section 3.1.4 will examine several additional devices that require extensions to the baseline process.

NOTE: The dimensions of standard bipolar devices are often specified in mils. A *mil* equals 0.001 inches. The relevant conversion factors are 1mil = 25.4μm, and $1\text{mil}^2 \cong 645\mu\text{m}^2$. Another obscure unit of measurement sometimes used to specify junction depths is the *sodium line*, which equals one-half of the wavelength of the sodium spectrum D-line (1 line = 0.295μm).[5]

NPN Transistors

Figure 3.10 shows a representative layout and cross section of a minimum-area NPN transistor. The collector of the NPN consists of an N-epi tank, and the base and emitter are fabricated by successive counterdopings. The carriers flow vertically from emitter to collector through the thin base region underneath the emitter diffusion. The difference between the base and emitter junction depths determines the effective base width. Since these dimensions are controlled entirely by diffusion processing, they are not subject to photolithographic misalignment, allowing a base width substantially smaller than the alignment tolerance. For example, a process with a 5μm minimum feature size can easily fabricate a 2μm base width.

---

[4] For a general overview of the standard bipolar process, see N. Doyle, "LIC Technology," *Microelectronics and Reliability*, Vol. 13, 1974, pp. 315–324.

[5] G. E. Anner, *Planar Processing Primer* (New York: Van Nostrand Reinhold, 1990), pp. 107–108.

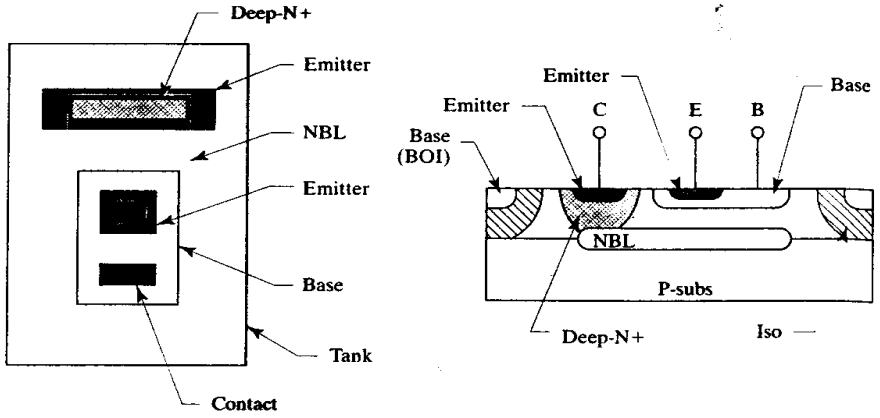


FIGURE 3.10 Layout and cross section of an NPN transistor with deep-N+ and NBL.[6]

The collector consists of lightly doped N-type epi lying on top of heavily doped NBL. The lightly doped epi allows the formation of a wide collector-base depletion region without excessive intrusion into the neutral base. This enables the transistor to support high operating voltages while simultaneously minimizing the Early effect (Section 8.2). NBL and deep-N+ create a low-resistance pathway to the portion of the epi layer beneath the transistor's active base. By this means, the collector resistance of a minimum NPN can be reduced to less than 100Ω and the collector resistance of a power NPN can be reduced to less than 1Ω.

The high concentration of donors in the NBL effectively halts the downward growth of the collector-base depletion region. The distance between the bottom of the base diffusion and the top of the NBL determines the maximum operating voltage of the NPN transistor. Thicker epi layers allow higher operating voltages, up to a limit set by the breakdown of the base diffusion sidewall (typically 50 to 80V). The maximum operating voltage of a bipolar process is usually specified in terms of the avalanche voltage of the NPN collector to emitter with base open (V_{CEO}). Depending on epi thickness and doping, this voltage can range from less than 10V to more than 100V.

The vertical NPN is the best device fabricated in the standard bipolar process. It consumes relatively little area and offers reasonably good performance. Circuit designers try to use as many of these transistors as possible. Table 3.1 lists typical device parameters for a minimum-emitter NPN transistor in a 40V standard bipolar process.

TABLE 3.1 Typical vertical NPN device parameters.

Parameter	Nominal Value
Drawn emitter area	$100\mu m^2$
Peak current gain (beta)	150
Early voltage	120V
Collector resistance, in saturation	100Ω
Collector current range for maximum beta	5μA–2mA
V_{EBO} (Emitter-base breakdown, collector open)	7V
V_{CBO} (Collector-base breakdown, emitter open)	60V
V_{CEO} (Collector-emitter breakdown, base open)	45V

[6] In many standard bipolar processes, the isolation spacings are much wider than this illustration suggests; see Appendix C.1 for a brief discussion and some typical spacing rules.

The NPN transistor can also act as a diode, the characteristics of which depend upon the terminals chosen to form the anode and cathode. The least series resistance and fastest switching speeds occur when the base and collector form the anode and the emitter forms the cathode. This configuration is sometimes called a *CB-shorted diode,* or a *diode-connected transistor.* Its only serious drawback consists of a low breakdown voltage, equal to the V_{EBO} of the transistor, or about 7V. On the other hand, the relatively low V_{EBO} allows a suitably-connected transistor to serve as a useful Zener diode. The breakdown voltage of this structure varies somewhat due to doping variations and surface effects, so a tolerance of at least ± 0.3V should be allowed.

PNP Transistors

The standard bipolar process cannot fabricate an isolated vertical PNP because it lacks a P-type tank. A non-isolated vertical PNP transistor, called a *substrate PNP,* can be constructed using the substrate as a collector. The collector of this device always connects to the substrate potential of the die, which usually consists of either ground or the negative supply rail. Figure 3.11 shows a representative layout and cross section of this device.

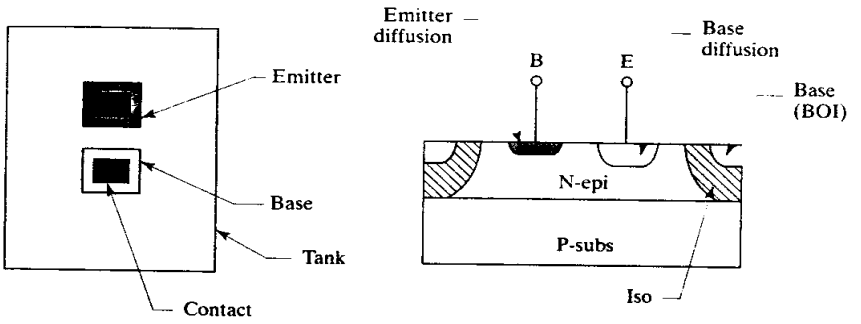


FIGURE 3.11 Layout and cross section of a substrate PNP transistor. The substrate forms the collector and is contacted through substrate contacts (not shown).

The base of the substrate PNP consists of an N-tank, and the emitter is fabricated from base diffusion. The collector current must exit through the substrate and the isolation. The collector contact does not have to reside next to the substrate PNP since all isolation regions interconnect electrically through the substrate. The resistance of the isolation and substrate are, however, substantial. Substrate contacts placed adjacent to the transistor help extract the collector current and thus minimize voltage drops in the substrate (*substrate debiasing*) that might otherwise impair circuit performance (Section 4.4.1).

The difference between the final epi thickness and the base junction depth determines the base width of a substrate PNP. As in the case of the vertical NPN, the base width is unaffected by photolithographic tolerances. NBL must be left out of the substrate transistor because its presence severely reduces beta. Deep-N+ therefore serves no useful function in a substrate PNP. An emitter diffusion placed beneath the collector contact ensures the surface doping concentration necessary to achieve Ohmic contact, while simultaneously thinning the oxide. The epi thicknesses and doping concentrations of the standard bipolar process are calculated to optimize the vertical NPN transistor, but the substrate PNP performs respectably (Table 3.2). The choice of names for the emitter and base diffusions is somewhat unfortunate, as the *emitter* of a substrate PNP consists of *base* diffusion.

TABLE 3.2 Typical PNP device parameters.

Parameter	Lateral PNP	Substrate PNP
Drawn emitter area	100μm²	100μm²
Drawn base width	10μm	N/A
Peak current gain (beta)	50	100
Early voltage	100V	120V
Typical operating current for maximum beta	5–100μA	5–200μA
V_{EBO} (Emitter-base breakdown, collector open)	60V	60V
V_{CBO} (Collector-base breakdown, emitter open)	60V	60V
V_{CEO} (Collector-emitter breakdown, base open)	45V	45V

The lack of an isolated collector limits the versatility of the substrate PNP. Another transistor, called a *lateral PNP*, trades off device performance for isolation. Figure 3.12 shows a representative layout and cross section of a minimum-geometry lateral PNP transistor. Both the collector and the emitter regions of the lateral PNP consist of base diffusions formed into an N-tank. As in the case of the substrate PNP, this tank serves as the base of the transistor. Transistor action in the lateral PNP occurs laterally outward from the central emitter to the surrounding collector. The separation of the two base diffusions sets the base width of the transistor. The emitter and collector of the lateral PNP are said to *self-align* because a single masking operation forms both regions. The base width of the lateral PNP can be precisely controlled because photolithographic misalignment does not occur between self-aligned diffusions. The effective base width of the transistor is considerably less than the drawn base width because of outdiffusion. This consideration limits the drawn base width to a minimum of about twice the base junction

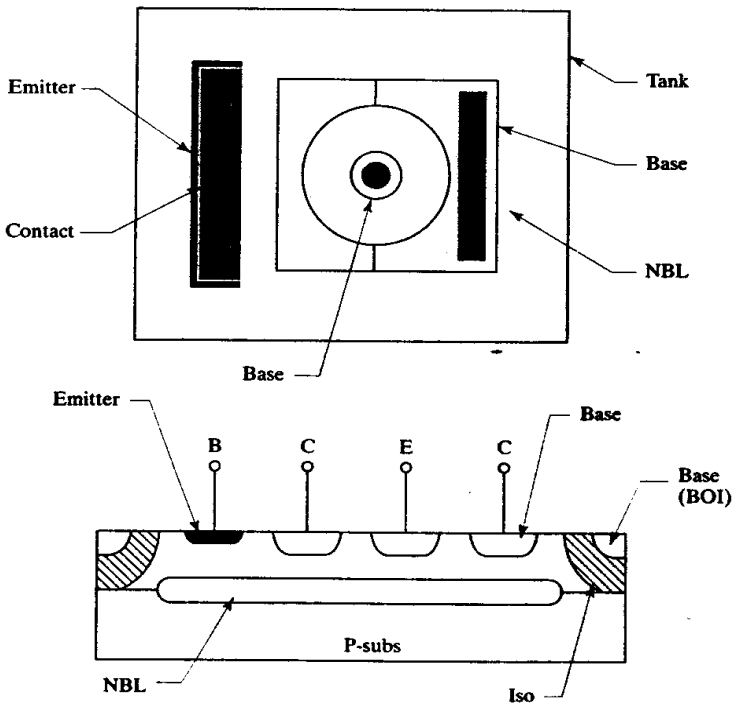


FIGURE 3.12 Layout and cross section of a lateral PNP transistor. The collector of the transistor appears twice on the cross section because it encircles the emitter.

depth. Narrow-base lateral PNP transistors exhibit low Early voltages and low-voltage punchthrough breakdown, so wider base widths are often employed.

Some percentage of the carriers injected by the lateral PNP's emitter will actually flow to the substrate rather than to the intended collector. This undesired conduction path forms a parasitic substrate PNP transistor. Unless this parasitic is somehow suppressed, most of the current injected by the emitter will find its way to the substrate, and the lateral PNP will exhibit very low apparent beta. For reasons that will be explained in Section 8.2.3, NBL largely blocks substrate injection and therefore boosts the lateral PNP beta.

Lateral PNP transistors have lower effective betas than their Gummel numbers would indicate. A large number of recombination centers reside at the oxide-silicon interface, especially in (111) silicon. The surface recombination rate thus far exceeds that in the bulk. Much of the current flow in the lateral PNP occurs near the surface and is therefore subject to these elevated recombination rates.[7] Certain layout techniques can minimize surface effects and. with care, betas of fifty or more can be obtained. Lateral PNP transistors are also quite slow, due mainly to large parasitic junction capacitances associated with the base terminal.

Neither the lateral nor the substrate PNP transistor forms a true complement to the vertical NPN. Both are useful devices. but each has its drawbacks and limitations. Circuit designers tend to avoid routing active signal paths through PNP devices (especially laterals) because of their poor frequency response, but most analog circuits still contain PNP transistors in supporting roles. Table 3.2 lists typical device parameters for PNP transistors formed on a 40V standard bipolar process.

Resistors

Standard bipolar does not include any diffusions intended specifically for fabricating resistors, but several types of resistors can be made using layers intended for other purposes. Typical examples include base, emitter, and pinch resistors, all three of which employ the relatively shallow base and emitter diffusions to achieve tighter spacings.

Each of the materials used to construct resistors possesses a characteristic *sheet resistance,* defined as the resistance measured across a square of the material contacted on opposite sides. Sheet resistance is customarily given in units of *Ohms per square* (Ω/□). It can be calculated from the thickness of the material and its doping concentration, but in the case of diffusions, nonuniform doping complicates these calculations (Section 5.2). In practice, sheet resistances are best determined by measuring sample resistors with known geometries constructed from the desired materials. Typical values for silicon diffusions range from 5 to 5000Ω/□.

A *base resistor* consists of a strip of base diffusion isolated by an N-tank and connected so that it will reverse-bias the base-epi junction (Figure 3.13). Connecting the tank to the more positive end of the resistor will ensure isolation. Alternatively, the tank can be connected to any point in the circuit biased to a higher voltage than the resistor. If a base resistor should forward-bias into its tank, a parasitic substrate PNP will inject current from the resistor into the substrate. NBL can help suppress this parasitic PNP should the base-epi junction momentarily forward-bias. Deep-N+ need not be added because the tank terminal does not draw significant current. Most standard bipolar processes produce base resistors with sheet resistances of 150 to 250Ω/□.

---

[7] R. S. Muller and T. I. Kamins. *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons. 1986). pp. 366–368.

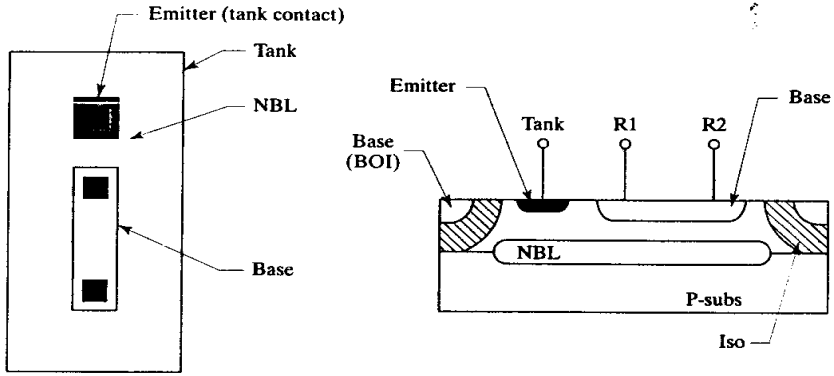


FIGURE 3.13 Layout and cross section of a base resistor.

An *emitter resistor* consists of a strip of emitter diffusion isolated by a base diffusion enclosed within an N-tank (Figure 3.14). The base region is connected so that it will reverse-bias the emitter-base junction, while the tank is biased to reverse-bias the base-epi junction. The simplest way to achieve this goal consists of tying the base to the low-voltage end of the resistor and the tank to the high-voltage end. Various other connections are also feasible, so long as neither junction forward-biases. NBL is usually added to help suppress parasitic substrate PNP action. The emitter sheet resistance is relatively low (typically less than 10Ω/□), and the breakdown of the emitter-base junction limits the differential voltage across the resistor to about 6V.

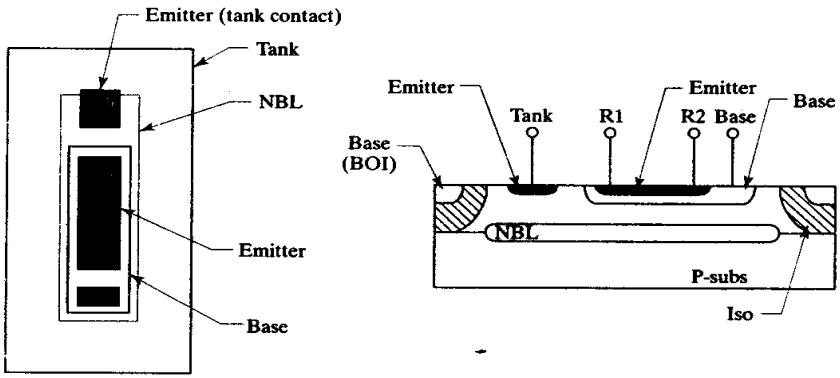


FIGURE 3.14 Layout and cross section of an emitter resistor (note the presence of bias terminals for both the tank and the base region).

A *pinch resistor* consists of a combination of base and emitter diffusions (Figure 3.15). The emitter forms a plate overlapping the middle of a thin strip of base.[8] Contacts occupy the ends of the base strip, which project from under the emitter plate. The tank and emitter plate are both N-type and are therefore electrically united. A tank contact biases both to a voltage slightly more positive than the resistor in order to ensure isolation. The body of the resistor consists of the portion of the base diffusion beneath the emitter plate. This *pinched base* is thin

[8] R. P. O'Grady, "The 'Pinch' Resistor in Integrated Circuits," *Microelectronics and Reliability*, Vol. 7, 1968. pp. 233–236.

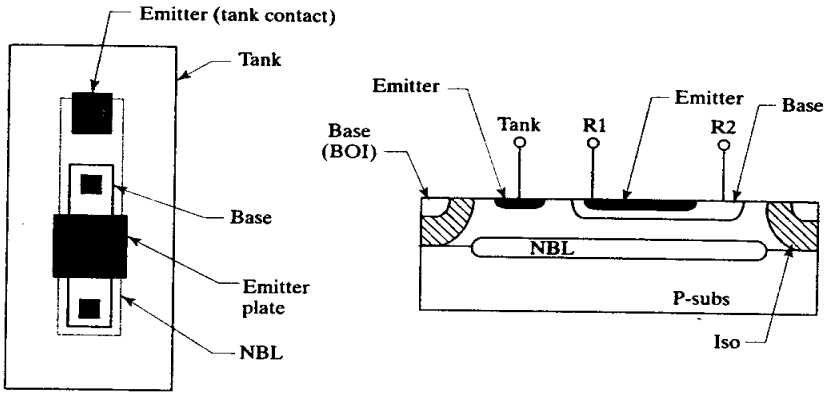


FIGURE 3.15 Layout and cross section of a base pinch resistor.

and lightly doped and therefore its resistance may exceed 5000Ω/□. Emitter-base breakdown limits the differential voltage across the resistor to about 7V. Pinch resistors are notoriously variable—much more so than either emitter or base resistors. Worst of all, these resistors exhibit severe voltage modulation. The intrusion of depletion regions into the neutral base tends to further pinch the resistor, causing it to act much like a JFET (Section 12.3). Pinch resistors find application in startup circuits and other noncritical roles, but their many drawbacks prohibit more widespread use. Table 3.3 compares the performance of emitter, base, and pinch resistors.

Parameter	Emitter	Base	Pinch
Sheet resistance	5Ω/□	150Ω/□	3000Ω/□
Minimum drawn width	8μm	8μm	8μm
Breakdown voltage	7V	50V	7V
Variability (15μm width)	±20%	±20%	±50% or more

TABLE 3.3 Typical resistor device parameters.

Capacitors

Standard bipolar was not intended to support capacitors. All of its oxide layers are so thick that they cannot be used to fabricate any but the smallest capacitors. However, the depletion region of a base-emitter junction exhibits a capacitance on the order of 0.8fF/μm^2 (0.5pF/mil^2), which can be used to construct a so-called *junction capacitor* (Figure 3.16). This capacitor consists of a base diffusion overlapping an emitter diffusion, both placed in a common tank. The emitter diffusion shorts to the tank, and the base-tank capacitance therefore adds to the base-emitter capacitance. The emitter plate must be biased positively with respect to the base plate to maintain a reverse bias across the base-emitter junction, and the differential voltage across the capacitor must not exceed the emitter-base breakdown voltage (about 7V). The resulting capacitance depends on bias and varies substantially (±50% or more). Junction capacitors are frequently used for compensating feedback loops, where their high capacitance per unit area makes up for their excessive variability.

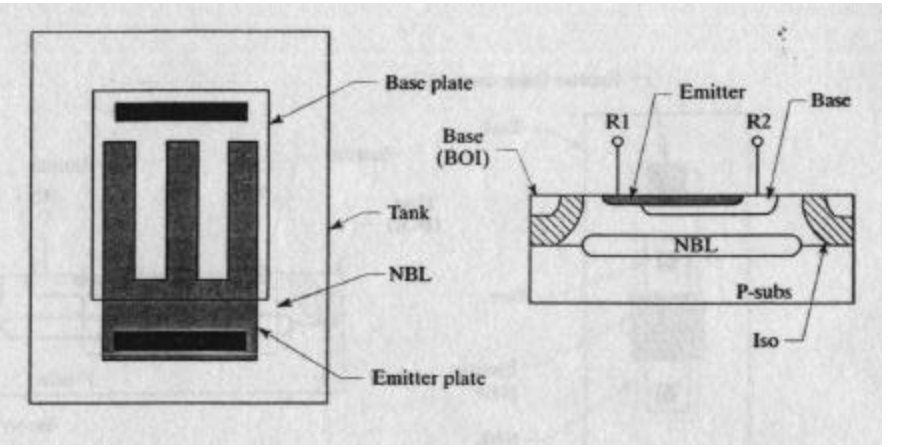


FIGURE 3.16 Layout and cross section of a junction capacitor.

3.1.4. Process Extensions

Standard bipolar has spawned a large number of process extensions, five of which are discussed in this section. They include up-down isolation, double-level metal, Schottky diodes, high-sheet resistors, and super-beta transistors. The benefits of each extension must be weighed against the increased cost and process complexity.

Up-down Isolation

Standard bipolar employs deep-P+ junction isolation driven down through the epitaxial layer to the underlying substrate. Outdiffusion increases the width of the isolation by 20μm or more, limiting how closely components can be packed together. One means of reducing outdiffusion uses a *P-buried layer* (PBL) to supplement the P+ isolation. The resulting *up-down isolation* consists of an isolation diffusion drawn coincident with the PBL. The isolation diffuses down from the surface, while the PBL diffuses up from the epi-substrate interface (Figure 3.17). Each only has to cross half the distance of a regular isolation diffusion, and outdiffusion is therefore cut approximately in half.

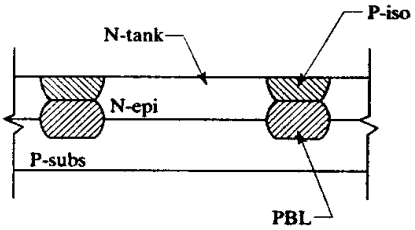


FIGURE 3.17 Cross section of a typical up-down isolation system.

Up-down isolation does have one significant drawback. Process considerations limit the PBL implant dose, so the final PBL becomes very lightly doped, and vertical resistance through the up-down isolation greatly exceeds that of conventional top-down isolation. The PBL also requires an additional masking step and diffusion. Up-down isolation saves 15 to 20% die area, so a case-by-case analysis should be performed to determine if it is worthwhile.

Double-level Metal

Standard bipolar originated as a *single-level metal* (SLM) process. The lack of a second metal layer greatly complicated lead routing. Instead of crossing wires by

means of jumpers, diffusions were employed to form low-value resistors, called *crossunders* or *tunnels* (Section 13.3.2). Many devices can be custom-tailored to incorporate crossunders at the cost of compromising device performance and increasing die area. Single-level routing requires a deep understanding of device and circuit operation and an intuitive sense of topological connectivity. Most layout designers require years to master these skills.

Double-level metal (DLM) can be added to a standard bipolar process at the cost of two extra masks: vias and metal-2. The thickness of the first metal layer is often reduced to simplify planarization. Double-level metal is a useful, if somewhat costly, option. Lead routing no longer requires the use of customized devices, allowing component standardization and a considerable reduction of layout time and effort. Since metallization consumes a great deal of area, double-level metal can also reduce die area by up to 30%. These benefits are so attractive that manufacturers now routinely employ double-level metal for all new designs.

Schottky Diodes

Standard bipolar originally used silicon-doped aluminum metallization. Modern processes usually employ a combination of silicidation and refractory barrier metallization to ensure adequate step coverage while maintaining low contact resistance. Along with its more obvious benefits, silicidation also offers the opportunity to fabricate reliable Schottky diodes. Although aluminum forms a rectifying Schottky barrier to lightly doped N-type silicon, the forward voltage of the resulting diodes varies unpredictably depending on annealing conditions. Certain silicides, most notably those of platinum and palladium, produce Schottky barriers with very stable and repeatable properties. The forward voltages of these Schottky diodes lie slightly below that of a moderately doped PN-junction diode, so they can serve as antisaturation clamps (Section 8.1.4).

Schottky diodes require contacts formed through thick-field oxide. A contact etch that just penetrates the field oxide will severely overetch the base and emitter contacts. Overetching can be prevented by performing two consecutive oxide removals, the first of which thins the field oxide over the Schottky contacts and the second of which creates the actual contact openings. The fabrication of Schottky diodes thus requires an additional masking step.

Figure 3.18 shows a typical Schottky diode layout. The anode consists of a rectifying contact to an N-epi tank while the cathode Ohmically contacts the same tank with the assistance of emitter diffusion. The addition of NBL and deep-N+ to the cathode

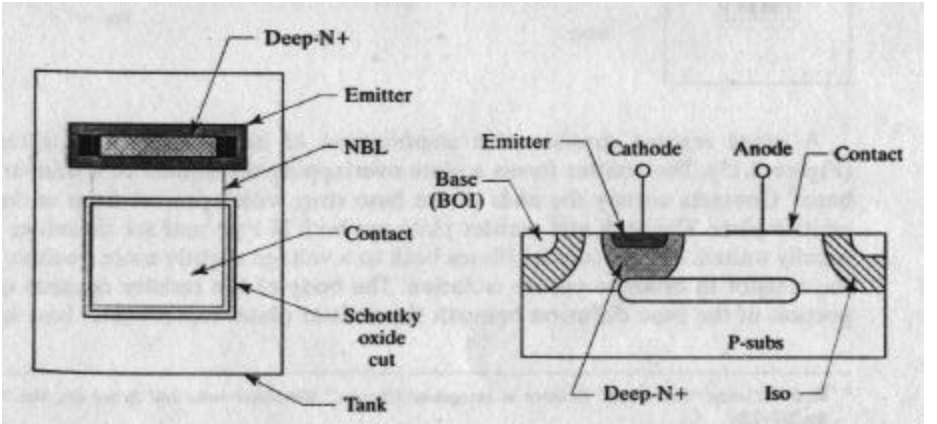


FIGURE 3.18 Layout and cross section of a Schottky diode.

greatly reduces the series resistance of the diode. The anode employs two concentric oxide removals, the larger of which thins the field oxide and the smaller of which forms the actual contact. This two-stage process not only eliminates overetching of base and emitter contacts but also improves the step coverage of metallization to the Schottky contact. This structure readily scales to provide diodes of any size. Electric field intensification at the exposed edges of the Schottky contact opening limits this simple structure's breakdown voltage and causes excessive reverse-bias leakage. Section 10.1.3 discusses methods of circumventing these problems.

High Sheet Resistors

Accurate resistors in standard bipolar are usually fabricated from base diffusion. Because the sheet resistance of this material rarely exceeds 200Ω/□, a typical die can incorporate a total of only 200 to 500kΩ of base resistance. Low-current circuits require more resistance than base diffusion can provide and more precision than pinch resistors can achieve. The only solution to this dilemma consists of adding a process extension to fabricate a precisely controlled high-sheet resistance (HSR) material.

The *high-sheet implant* consists of a shallow, lightly doped P-type implant. Depending on dose and junction depth, the sheet resistance of this implant can range from 1 to 10kΩ/□. The larger sheet values suffer from surface depletion effects that cause resistance variations, so most processes employ HSR implants of 1 to 3kΩ/□.

Figure 3.19 shows the layout and cross section of a typical HSR resistor. The body of the resistor consists of high-sheet implant, but small regions of base diffusion at either end of the resistor ensure Ohmic contact. The resistor occupies an N-tank and is isolated by the reverse-biased HSR-tank junction. The tank is often connected to the more positive end of the resistor, just as in the case of the base resistor. High-sheet resistors require one additional mask step and one dedicated implant. The cost of this process extension can usually be justified if the circuit includes more than 100 to 200kΩ of resistance.

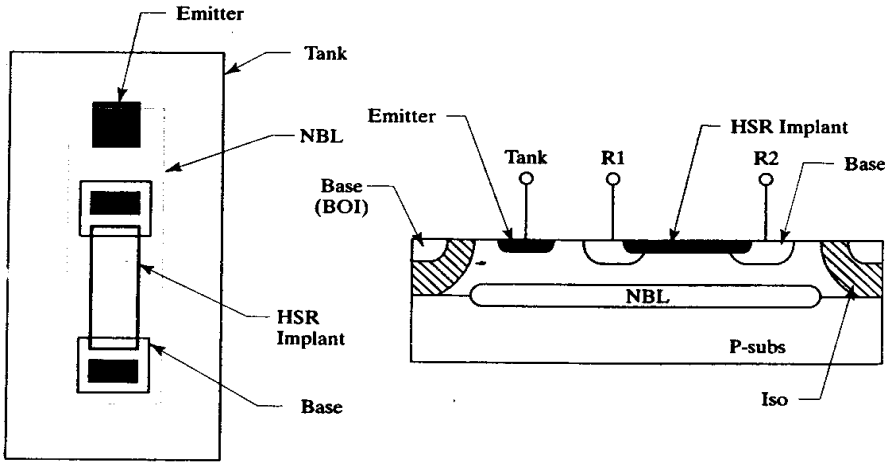


FIGURE 3.19 Layout and cross section of a high-sheet resistor.

Super-beta Transistors

The standard bipolar NPN provides a reasonable compromise between high beta and adequate operating voltage. Beta can be greatly increased by narrowing the base width. If the process incorporates two separate emitter implants, then one can

be optimized for beta and the other for operating voltage. The resulting *super-beta transistors* can achieve betas of 1000 to 3000 (Section 8.3.1). The use of an extremely thin and lightly doped base causes considerable depletion region intrusion into the neutral base and hence compromises not only operating voltage but also Early voltage. These highly specialized devices find application only in a limited range of circuits. For example, they have been used to fabricate bipolar operational amplifiers with extremely low input bias currents.

3.2 POLYSILICON-GATE CMOS

With the addition of two masking steps, standard bipolar can fabricate metal-gate PMOS transistors similar to those fabricated by early MOS processes (Figure 3.20). An N-tank serves as a backgate for the PMOS transistor; the backgate contact incorporates emitter diffusion to ensure Ohmic contact. None of the oxide layers in standard bipolar is sufficiently thin to serve as a gate oxide, necessitating the addition of a special masking step. Aluminum metal forms the gate electrode, while the source and drain consist of shallow P+ implants. Since standard bipolar does not include any suitable implant, another masking step patterns a special P-type source/drain (PSD) implant.

Practical MOS transistors require threshold voltages that lie within relatively narrow limits. Threshold voltages of less than 0.5V cause excessive leakage, while those of more than 1.5V unnecessarily reduce the available transconductance. The threshold voltage of an unadjusted (or *natural*) PMOS usually lies between 2 and 4V, necessitating a threshold adjust implant to shift it to the desired target of about 1V. The threshold voltage must lie within approximately ±0.5V of target. The metal gate PMOS transistor has difficulty maintaining even this minimal degree of control. The excess surface state charges introduced by using (111) silicon constitute one source of threshold variation; mobile ion contamination (Section 4.2.2) represents another.

Metal-gate MOS transistors also suffer from excessive overlap capacitance. The gate electrode is patterned by a different mask than the source and drain diffusions are. The gate must therefore overlap the source and drain sufficiently to form a continuous channel, even in the presence of photolithographic misalignments. The overlap between the gate and source causes a gate-to-source capacitance C_{gs}, while the

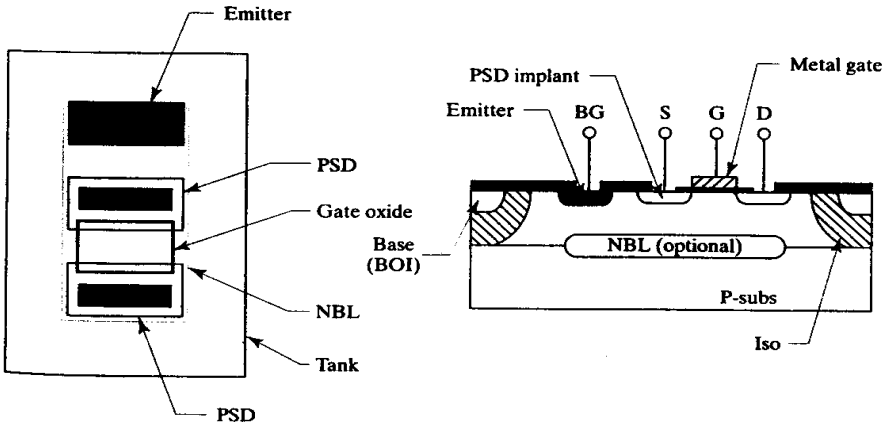


FIGURE 3.20 Layout and cross section of a metal-gate PMOS transistor constructed in standard bipolar.

overlap between gate and drain causes a gate-to-drain capacitance C_{gd}. These parasitic capacitances slow the transistor because they must be charged and discharged during switching. The gate-to-drain capacitance is particularly deleterious because the voltage gain of the transistor multiplies its apparent value (a phenomenon called the *Miller effect*). These parasitic overlap capacitances must be minimized in order to allow the construction of high-speed logic circuitry.

A complementary NMOS transistor would greatly enhance the utility of the metal-gate PMOS process. Taken together, NMOS and PMOS transistors would allow the construction of versatile *complementary MOS* (CMOS) circuits. Unfortunately, standard bipolar cannot easily fabricate NMOS transistors because they require a lightly doped P-type backgate that does not exist in this process. The NMOS threshold voltage on suitably doped (111) silicon is slightly negative, so a threshold-adjust implant is required to form an enhancement device. Yet another masking step is required to raise the thick-field threshold in the lightly doped P-type backgate around the NMOS transistors to prevent parasitic channel formation. The relatively poor performance of metal-gate CMOS cannot justify the cost of five additional masking steps, especially when a nine-mask polysilicon-gate CMOS process can fabricate vastly superior transistors. The following section examines the construction and performance of this alternate process.

3.2.1. Essential Features

The polysilicon-gate CMOS process is optimized to form complementary PMOS and NMOS transistors on a common substrate. It does not support the construction of bipolar transistors and it offers only a limited range of passive components. Originally intended solely for manufacturing CMOS logic gates, with slight modifications this process can also fabricate a limited variety of analog circuits.

A key difference between polysilicon-gate CMOS and standard bipolar lies in the choice of substrate material. Standard bipolar employs (111) silicon to enhance the thick-field threshold by increasing the surface state density, while polysilicon-gate CMOS uses (100) silicon to reduce the surface state density in order to improve threshold voltage control. A second major innovation lies in the use of polysilicon rather than aluminum as the gate material. Polysilicon can safely withstand the high temperatures required to anneal the source/drain implants, so it can act as a mask to form self-aligned sources and drains. The effects of mobile ion contamination can also be minimized by doping the polysilicon gate with phosphorus. Poly gates thus offer not only faster switching speeds but also better control of threshold voltages.

The choice of threshold voltages forms one of the few differences between analog and digital CMOS processes. Most digital CMOS processes[9] target threshold voltages between 0.8V and 0.9V with a variation of about ±0.2V. Analog CMOS designers favor maximizing headroom by targeting threshold voltages around 0.7V ± 0.2V. Since (100) silicon dictates the use of threshold-adjust implants in either case, the threshold voltages can often be retargeted by simply changing the implant dosage. Analog CMOS also rules out the use of blanket silicidation (Section 3.2.4). Neither of these requirements fundamentally modify the polysilicon-gate CMOS process.

[9] The information provided in the text applies to processes with operating voltages of 5V or more. Lower-voltage processes require smaller threshold voltages and tighter control. For example, a 3V process will typically target a threshold voltage of 0.6 ± 0.15V.

3.2.2. Fabrication Sequence

The baseline polysilicon-gate CMOS fabrication sequence consists of nine masking operations. The processing steps required to fabricate a finished wafer will be presented in the order in which they are performed. The cross sections used to illustrate this process employ a vertical exaggeration of between two and five, just as did the cross sections previously presented for standard bipolar.

Starting Material

CMOS integrated circuits are normally fabricated on a P-type (100) substrate doped with as much boron as possible in order to minimize substrate resistivity. This precaution helps provide a degree of immunity to *CMOS latchup* by minimizing substrate debiasing (Section 4.4). CMOS processes do not require NBL, so substrate doping is limited only by solid solubility.

Epitaxial Growth

The first step of the CMOS process consists of growing a lightly doped P-type epitaxial layer on the substrate. This epitaxial layer, typically some 5 to 10μm thick, is considerably thinner than the one used for standard bipolar. NMOS transistors are formed directly into the epi layer, which serves as their backgate. Since this process needs no buried layers, epi-coated wafers can be stockpiled to serve as starting material for all types of products. Standard bipolar does not allow this economy of scale, since each product requires a uniquely patterned NBL.

In theory, CMOS processes do not require epitaxy since the MOS transistors can be grown directly into a P– substrate. Epitaxial processing increases costs, but it also improves latchup immunity by allowing the use of a P+ substrate. In addition, the electrical properties of the epitaxial layer are more precisely controllable than those of Czochralski silicon, resulting in better control of MOS transistor parameters.

N-well Diffusion

After the wafer has been thermally oxidized, a layer of photoresist that has been spun onto it is patterned using the N-well mask. An oxide etch opens windows through which ion implantation deposits a controlled dose of phosphorus. A prolonged high-temperature drive creates a deep lightly doped N-type region called an *N-well* (Figure 3.21). The N-well for a typical 20V CMOS process has a junction depth of about 5μm. Thermal oxidation during the well drive covers the exposed silicon with a thin layer of *pad oxide.*

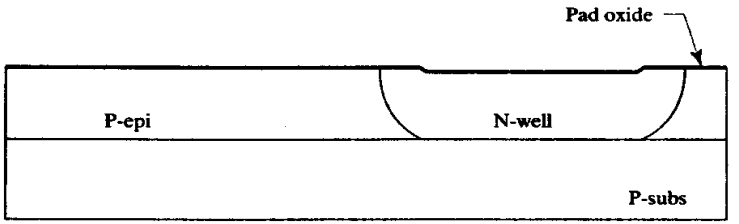


FIGURE 3.21 Wafer after N-well drive.

In an *N-well CMOS process,* such as that illustrated in Figure 3.21, NMOS transistors occupy the epi, and PMOS transistors reside in the well. The increased total dopant concentration caused by counterdoping the well slightly degrades the mobility of majority carriers within it. The N-well process therefore optimizes the performance of the NMOS transistor at the expense of the PMOS transistor. As a side

effect, the N-well process also produces the grounded substrate favored by most circuit designers.

A *P-well CMOS process* uses an N+ substrate, an N– epitaxial layer, and a P-well. NMOS transistors are formed in the P-well and PMOS transistors in the epi. This process optimizes the PMOS transistor at the expense of the NMOS transistor, but the NMOS still outperforms its counterpart because electrons are more mobile than holes. A P-well process requires that the substrate connect to the highest-voltage supply instead of ground. Designs that employ multiple power supplies often have difficulty biasing an N-type substrate because of ambiguities in the sequencing of the supplies.

Both P-well and N-well CMOS processes exist. The N-well process offers a slightly better NMOS transistor, and it allows the use of a grounded substrate. N-well CMOS is also upwardly compatible with BiCMOS technology, as will become apparent later in this chapter. The N-well process has therefore been chosen to illustrate CMOS technology.

Inverse Moat

The CMOS process employs a thick-field oxide for much the same reasons as standard bipolar: it increases the thick-field threshold voltages, and it reduces parasitic capacitance between the metallization and the underlying silicon. Unlike standard bipolar, CMOS processes employ LOCOS technology to selectively grow the field oxide, leaving only a thin pad oxide over the regions where active devices will be formed. The locally oxidized regions of the die are called *field regions*, while the areas protected from oxidation are called *moat regions*.

The LOCOS process uses a patterned nitride layer formed by first depositing nitride across the entire wafer, then patterning this nitride using the inverse moat mask, and finally employing a selective etch to remove the nitride over the field regions (Figure 3.22). The photomask used for this step is called the *inverse moat* mask because it consists of a color reverse of the moat regions. In other words, the mask codes for areas where moat is absent, not where it is present.

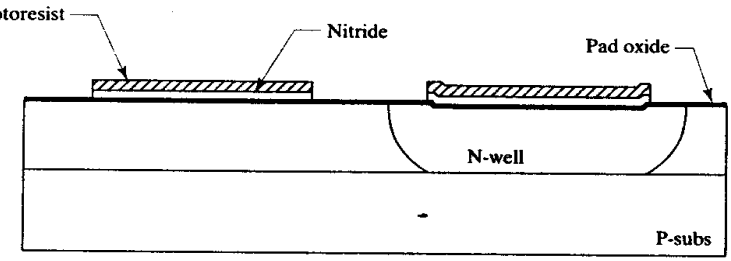


FIGURE 3.22 Wafer after nitride deposition and inverse moat pattern.

The nitride layer used for LOCOS must lie on top of a thin oxide layer (the *pad oxide*) because the conditions of nitride growth induce mechanical stresses that can cause dislocations in the silicon lattice. The pad oxide provides mechanical compliance and prevents strains produced by the nitride growth from damaging the underlying silicon.

Channel Stop Implants

The CMOS process deliberately minimizes threshold voltages in order to produce practical MOS transistors. The LOCOS field oxide will raise the thick-field thresholds, but not by enough to support supply voltages of more than a few volts. Practical CMOS processes nearly always require additional measures to ensure that

the thick field thresholds exceed the operating voltages. Dopants are usually selectively implanted beneath the field regions to raise the threshold voltages of the thick-field transistors. P-epi field regions receive a P-type *channel stop implant,* while N-well field regions receive an N-type channel stop implant. The formation of channel stops thus requires two successive ion implantations.

Various techniques have been developed to produce channel stops. The method presented here involves the use of a blanket boron implant followed by a patterned phosphorus implant. The boron implant uses the photoresist left from patterning the LOCOS nitride. This mask exposes the field regions where channel stops will be deposited, so all of these regions receive the blanket boron implant (Figure 3.23A). This step sets the thick-field threshold in the epi regions.

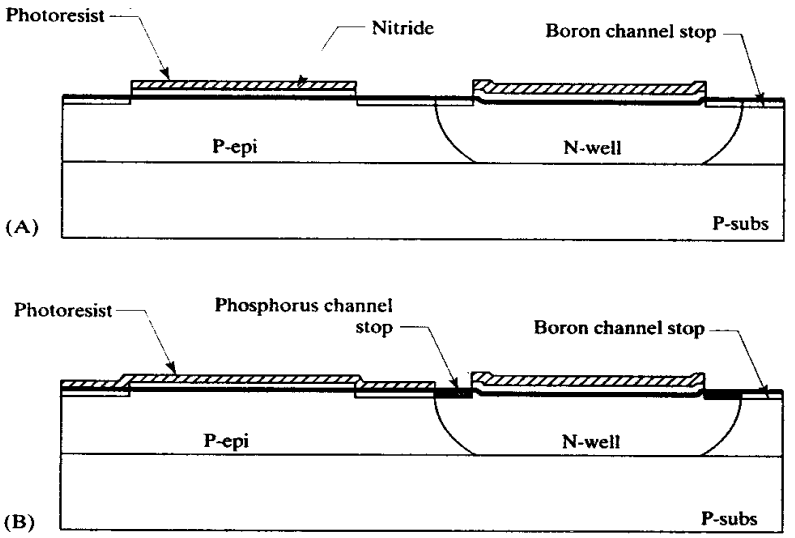


FIGURE 3.23 Wafer after blanket boron channel stop implant (A) and after selective phosphorus channel stop implant (B).

The wafer is again coated with photoresist immediately after the boron implant. The previous photoresist can remain in place since the channel stop implant will not affect the moat regions that lie beneath it. The recoated wafer is patterned using the channel stop mask, exposing only N-well field regions. The subsequent phosphorus implant counterdopes the previous blanket boron implant and raises the NMOS thick-field threshold above the maximum operating voltage (Figure 3.23B). Following the phosphorus implant, all photoresist is stripped from the wafer in preparation for LOCOS oxidation.

LOCOS Processing and Dummy Gate Oxidation

Steam is often used to increase the rate of LOCOS oxidation; alternately the furnace pressure can be raised to five or ten times atmospheric. After LOCOS oxidation, a suitable etchant strips away the remnants of the nitride block mask. Figure 3.24 shows a cross section of the resulting wafer. The curved transition region, called a *bird's-beak,* at the edges of the moat results from oxidants diffusing under the edges of the nitride film.

The *Kooi effect* (Section 2.3.4) causes nitride deposits to form underneath the pad oxide around the edges of the moat. These deposits can potentially cause gate oxide integrity failures, but they can be eliminated by a dummy gate oxidation. A

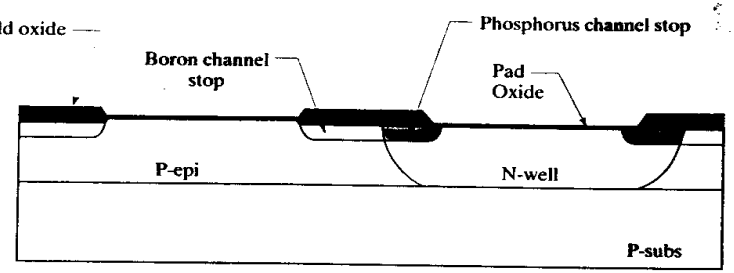


FIGURE 3.24 Wafer after LOCOS oxidation and nitride strip.

brief etch strips away the thin pad oxide without substantially eroding the thick-field oxide. Next, a brief dry oxidation grows a thin layer of oxide called a *dummy gate oxide* (or *sacrificial gate oxide*) in the moat regions. Any nitride deposits that remain will gradually oxidize. All of the nitride will be consumed if the dummy gate oxidation continues for a sufficient length of time.

Threshold Adjust

The use of (100) silicon helps stabilize the threshold voltages of the MOS transistors, but the backgate dopings and gate electrode materials preclude the achievement of usable threshold voltages without threshold adjust implants. For example, the unadjusted PMOS threshold might range from –1.5 to –1.9V, while the NMOS threshold might range from –0.2 to 0.2V. One or two threshold adjust implants (also known as V_t *adjusts*) retarget the threshold voltages to the desired targets, usually 0.7V for NMOS and –0.7V for PMOS transistors.

Two methods exist for adjusting threshold voltages. The first method employs two separate implants, one to set the PMOS V_t and the other the NMOS V_t. The use of two implants allows independent optimization of both thresholds. Many processes do not require this degree of flexibility. These processes can use a single V_t adjust to simultaneously reduce the PMOS threshold and increase the NMOS threshold. If this implant is properly performed, a nominal threshold voltage of 0.7 to 0.9V can be obtained for both types of MOS transistors. Figure 3.25 illustrates this approach.

After the wafer has been coated with photoresist, the V_t adjust mask is used to open windows over areas where MOS transistors will form. The boron V_t adjust implant penetrates the dummy gate oxide to dope the underlying silicon. After the V_t adjust implant, the dummy gate oxide is stripped away to reveal bare silicon in the moat regions.

The true gate oxidation employs dry oxygen to minimize excess charge incorporation due to surface states and fixed oxide charges. This oxidation must be very brief, because gate oxides are exceedingly thin. A 10V MOS transistor typically requires a 300Å (0.03μm) gate oxide, while a 3V transistor may employ an oxide less than 100Å (0.01μm) thick. This gate oxide will form the dielectric of the MOS

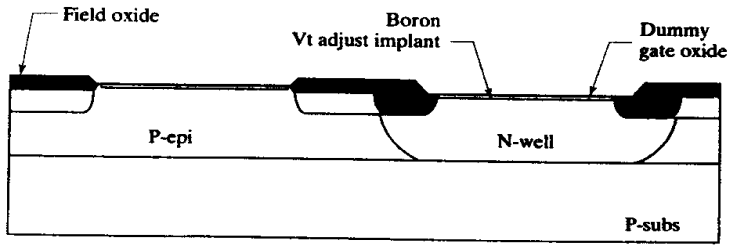


FIGURE 3.25 Wafer after V_t adjust implant.

transistors; it also covers the regions where source and drain implants will later occur.

Polysilicon Deposition and Patterning

The polysilicon layer used to form gate electrodes is heavily doped with phosphorus to reduce its resistivity to about 20 to 40Ω/□. Although gate leads do not conduct DC current, switching transients do draw substantial AC current, and low-resistance gate polysilicon greatly improves the switching speeds of MOS circuitry. Phosphorus doping simultaneously adjusts the work function of the polysilicon to produce threshold voltages compatible with a single-step V_t adjust. Phosphorus-doped gate polysilicon also minimizes threshold voltage variation due to mobile ions, allowing threshold voltage control of ±0.1 to 0.2V. While it is possible to dope polysilicon during deposition, most processes first deposit intrinsic polysilicon and subsequently dope it using conventional deposition or implantation techniques.

The deposited polysilicon layer must now be patterned using the poly mask (Figure 3.26). Modern submicron processes can fabricate polysilicon gates less than 0.5μm long, and any variation in gate length directly affects the transconductance of the resulting transistors. Thus, the patterning and etching of poly form the most critical photolithographic steps of a CMOS process. The simple process discussed here produces a minimum channel length of about 2μm and therefore does not require as high a degree of precision as submicron processes, but polysilicon patterning still remains its most challenging photolithographic step.

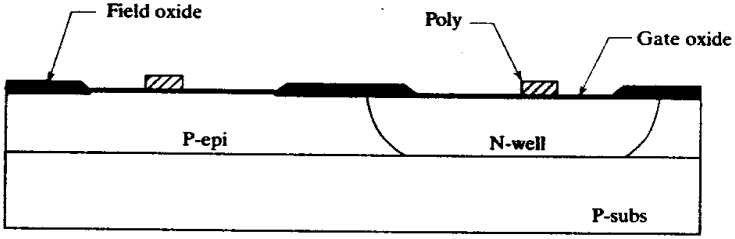


FIGURE 3.26 Wafer after polysilicon deposition and pattern. For simplicity, the channel stop and threshold adjust implants do not appear in this or subsequent cross sections.

Source/Drain Implants

The completed polysilicon gates now act as masks that self-align the source/drain implants for both the PMOS and NMOS transistors. These implants can be performed in either order. In the illustrated process, the N-type source/drain (NSD) implant occurs first, followed by the P-type source/drain (PSD) implant.

The NSD implant begins with the application of photoresist to the wafer, followed by patterning using the NSD mask. Shallow, heavily doped N-type regions are then formed by implanting arsenic through the exposed gate oxide. The polysilicon gate blocks this implant from the regions directly underneath the gate and therefore minimizes the gate/source and gate/drain overlap capacitances. Once the NSD implant has been completed, the photoresist residue is stripped from the wafer. The PSD implant begins with the application of a second photoresist layer patterned using the PSD mask. A shallow, heavily doped P-type region is formed by implanting boron through the exposed gate oxide. As with the NSD implant, the PSD implant self-aligns to the polysilicon and the PMOS transistors also exhibit minimal overlap capacitance. Following the PSD implant, photoresist is again stripped from the wafer.

A brief anneal activates the implanted dopants and slightly thickens the oxide over the source and drain regions. This anneal is the final high-temperature step of

the process, corresponding to the emitter drive of standard bipolar. Figure 3.27 shows a cross section of the wafer following the source/drain anneal.

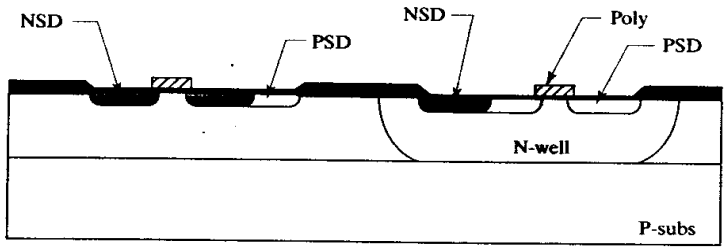


FIGURE 3.27 Cross section of the wafer after NSD and PSD implants and anneals. The backside contact implants abut the source implants to save space.

Contacts

Despite further oxidation during the source/drain anneal, the oxide covering the moat regions remains thin and therefore vulnerable to oxide rupture. Most processes deposit a *multilevel oxide* (MLO) before contact patterning. The MLO thickens the oxide over the moat regions and at the same time coats and insulates the exposed polysilicon pattern. Metal leads can now run over moat regions and polysilicon gates without risk of oxide rupture.

After the wafer is again coated with photoresist, the contact regions are patterned using the contact mask. Ohmic contacts form to the heavily doped source and drain without difficulty, but the backgate regions are far too lightly doped to allow direct Ohmic contact. The addition of NSD and PSD implants in the vicinity of the backgate contacts overcomes this difficulty. Contacts opened over polysilicon allow contact to the gate electrodes.

Metallization

The shallow NSD and PSD diffusions are vulnerable to junction spiking. Most CMOS processes employ a combination of contact silicidation and refractory barrier metallization to ensure reliable contact to the source/drain regions. After formation of silicide in the contact openings, a thin film of refractory metal sputtered over the wafer precedes a much thicker layer of copper-doped aluminum. The metallized wafer is coated with photoresist and patterned, using the metal mask. A suitable etchant then removes unwanted metal to form the interconnection pattern. Most processes also include a second layer of metallization. In such a process, another layer of oxide deposited over the first metal pattern insulates it from the second metal pattern. This second deposited oxide is usually called an *interlevel oxide* (ILO). Some form of planarization minimizes the nonplanarities caused by the first metal pattern to ensure adequate second metal step coverage. Vias etched through the ILO connect to a second metal layer deposited and patterned in much the same way as the first.

Protective Overcoat

A protective overcoat is now deposited over the final layer of metallization, both to provide mechanical protection and to prevent contamination of the die. The protective overcoat must resist penetration by mobile ions, so it normally consists of either a thick phosphosilicate glass (PSG), a compressive nitride layer, or both.

After coating with photoresist, the wafer is patterned using the *protective overcoat* (PO) mask. A suitable etchant removes the overcoat over selected areas of

metallization and allows attachment of bondwires to the integrated circuit. This composes the final fabrication step; the wafer is now complete. Figure 3.28 shows a cross section of the resulting wafer, with only a single level of metal for simplicity. No bondpad openings exist in the illustrated portion of the die. This cross section includes an NMOS transistor on the left and a PMOS transistor on the right.

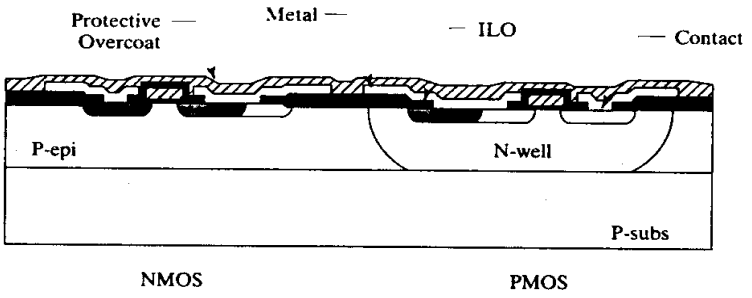


FIGURE 3.28 Cross section of the completed polysilicon-gate CMOS wafer.

3.2.3. Available Devices

Polysilicon-gate CMOS was originally developed to provide relatively low-voltage NMOS and PMOS transistors. The same process steps can fabricate several other components, including natural MOS transistors, a substrate PNP, several types of resistors, and a capacitor. Together these components allow the construction of a considerable variety of analog circuits. Section 3.2.4 examines process extensions that allow higher operating voltages and denser circuit packing.

NMOS Transistors

Figure 3.29 shows a representative layout and cross section of an NMOS transistor. The source and drain regions consist of NSD implants that self-align to the polysilicon gate. Since the backgate of the NMOS consists of the P-epi (and by extension the substrate), any substrate contact on the die will serve as a backgate terminal for the transistor. Many layouts actually include separate backgate contacts immediately adjacent to each NMOS transistor even though these are not strictly necessary. The close proximity of these backgate contacts improves CMOS latchup immunity, and this arrangement ensures that an adequate number of substrate contacts are distributed throughout the layout. In cases where the source of the NMOS transistor connects to the substrate potential, a very compact layout can be achieved by butting the PSD substrate contact against the NSD source. PSD and NSD cannot abut one another if they connect to different potentials because the resulting P+/N+ junction leaks excessively.

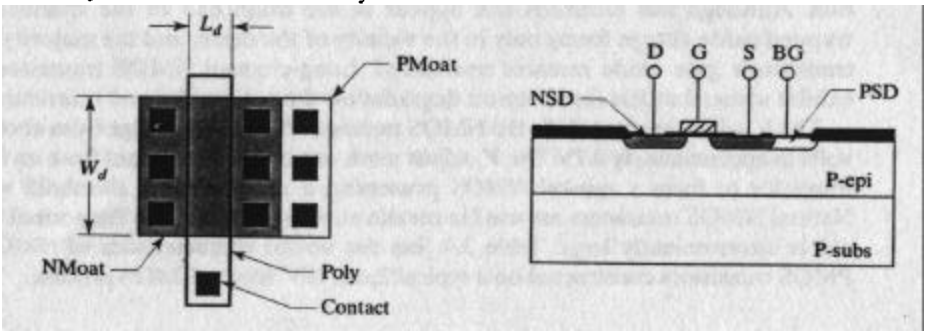


FIGURE 3.29 Layout and cross section of an NMOS transistor. The source and backgate of this transistor are shorted together using metal (not shown).

Figure 3.29 illustrates the common practice of coding CMOS transistors using drawing layers called *NMoat* and *PMoat*. These layers do not correspond to individual masks but rather to combinations of several masks. A figure drawn on the NMoat layer simultaneously produces figures on both the NSD and moat masks. Similarly, a figure drawn on PMoat simultaneously generates figures on both the PSD and moat masks. The use of PMoat and NMoat drawing layers simplifies the layout by reducing the number of figures required to draw a transistor.

The arrays of small square contacts shown in Figure 3.29 are characteristic of CMOS processes. The etch rate of oxide windows varies somewhat depending upon their size and shape, and these variations become particularly severe for very small openings. Many processes therefore allow only a single size of contact opening, usually consisting of a minimum-dimension square. Larger contacts must consist of arrays of minimum contacts rather than larger oxide openings.

In Figure 3.29, W_d and L_d denote the *drawn width* and *drawn length* of the NMOS transistor, respectively. The names given to these two dimensions may seem counterintuitive since the length of the gate is actually the width of the drawn polysilicon strip, but the channel length is customarily defined as the separation between its source and drain regions. The transconductance of an MOS transistor is approximately proportional to the ratio of the channel width divided by the channel length (W_d/L_d). Short channel lengths generate more transconductance per unit area, but analog circuits often employ longer channels to reduce channel length modulation.

Hot electron degradation limits the simple NMOS transistor of Figure 3.29 to relatively low operating voltages. The depletion region across the pinched-off portion of the channel accelerates electrons to high velocities. Some of these *hot electrons* collide with lattice atoms and ricochet out of the channel into the overlying gate oxide, where they become trapped. Hot electron injection causes a gradual shift in threshold voltage due to the slow accumulation of a fixed oxide charge. Eventually the threshold voltage shifts so far that the circuit ceases to meet parametric specifications.

Hot electron injection only occurs when the NMOS transistor operates in saturation with a relatively large drain-to-source bias. In the linear region the drain-to-source voltage is too small to produce hot electrons, and in cut-off no conduction occurs. NMOS transistors used as switches experience hot electron injection only during switching transients. If the switching frequency remains fairly low, then the total quantity of hot electrons produced over the operating lifetime of the integrated circuit remains acceptably small. Two different operating voltages are often specified for NMOS transistors. Junction breakdown and punchthrough limit the *blocking voltage rating*, which applies to transistors used as switches and as components of low-frequency digital logic. The somewhat lower *operating voltage rating* determined by the onset of hot electron degradation applies to transistors that operate for an appreciable length of time in saturation (as do the majority of analog transistors).

Increasing the length of the transistor reduces the impact of hot electron injection. Although hot electrons still appear at the drain end of the transistor, the trapped oxide charge forms only in the vicinity of the drain, and the majority of the transistor's gate oxide remains unaffected. Long-channel NMOS transistors thus exhibit somewhat less hot electron degradation than short-channel transistors.

The V_t adjust implant shifts the NMOS transistor threshold voltage from about zero volts to approximately 0.7V. The V_t adjust mask can block the implant from an NMOS transistor to form a *natural NMOS* possessing a relatively low threshold voltage. Natural NMOS transistors are used in certain analog circuits where the normal threshold is inconveniently large. Table 3.4 lists the device characteristics of NMOS and PMOS transistors constructed on a typical 2μm, 10V analog CMOS process.

Parameter	NMOS	PMOS
Minimum channel length	2μm	2μm
Gate oxide thickness	400Å	400Å
Threshold voltage (adjusted)	0.7V	–0.7V
Threshold voltage (natural)	0V	–1.4V
Transconductance (W_d/L_d = 10/10)	50μA/V^2	20μA/V^2
Operating voltage	7V	15V
Blocking voltage	15V	15V

TABLE 3.4 Typical polysilicon-gate CMOS device parameters.[10]

PMOS Transistors

Figure 3.30 shows a representative layout and cross section of a PMOS transistor. This device resides in an N-well that acts as its backgate. Any number of PMOS transistors can occupy the same well as long as their backgates all tie to the same potential. The relatively deep N-well outdiffuses substantially and the layout dimensions associated with it become quite large. Merging PMOS transistors in the same tank therefore saves substantial area. While the backgate of an NMOS transistor inherently connects to substrate, the backgate of a PMOS transistor can connect to any potential greater than or equal to its source. The N-well backgate thus provides an extra degree of freedom that analog designers frequently employ to enhance circuit performance.

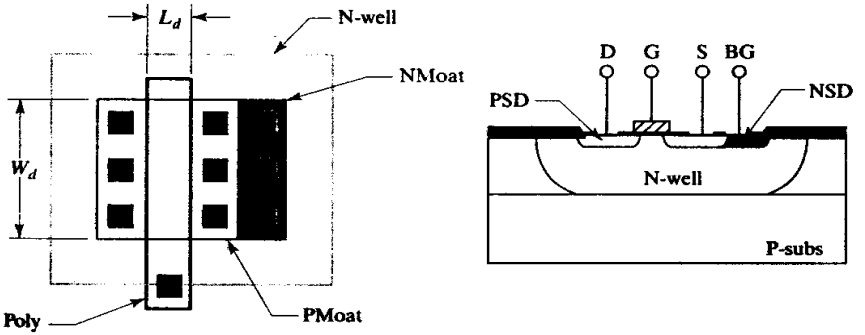


FIGURE 3.30 Layout and cross section of a PMOS transistor. The backgate and source of this transistor are shorted together using metal (not shown).

PMOS transistors are subject to *hot hole degradation,* but this causes fewer problems than hot electron degradation because holes are less mobile than electrons. A higher electric field and therefore a larger drain-to-source voltage is required to accelerate holes to velocities sufficient to inject charge into the oxide. Junction avalanche and punchthrough often limit PMOS transistors to voltages where hot hole degradation remains unimportant. Higher-voltage PMOS transistors will encounter hot-carrier problems similar to those of NMOS transistors.

A *natural PMOS* transistor can be fabricated by blocking the V_t adjust implant from the channel region of the device. Natural PMOS transistors possess inconveniently high threshold voltages, usually in excess of a volt. These transistors see only

[10] These parameters are roughly analogous to those of the 3μm Advanced LinCMOS™ process described in R. K. Hester, L. Hutter, L. Le Toumelin, J. Lin, and Y. Tung. "Linear CMOS Technology," *TI Technical Journal,* Vol. 8, #1, 1991, pp. 29–41. (The trademark belongs to Texas Instruments.) The transconductance figures given in this text are those used in the Schichman-Hodges equation $I_d = 1/2\ k\ (W/L)\ (V_{gs} - V_t)^2$.

occasional use in analog circuit design. Table 3.4 lists typical device parameters for a 2μm, 10V PMOS transistor.

Substrate PNP Transistors

The only bipolar transistor available in an N-well process is a substrate PNP. Figure 3.31 shows a typical layout and cross section for this device. The emitter consists of a PSD implant formed in an N-well that acts as the base of the transistor. A slug of NSD provides Ohmic contact to the N-well. The collector of this device consists of the P+ substrate and P-epi surrounding the well.

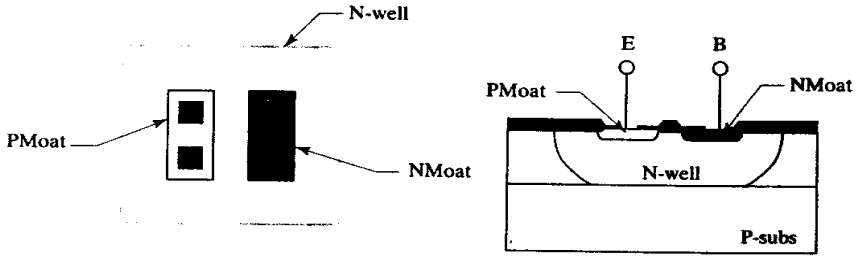


FIGURE 3.31 Layout and cross section of a substrate PNP transistor. The collector is connected by means of substrate contacts (not illustrated).

Although CMOS processes do not deliberately optimize bipolar components, the substrate PNP can still provide reasonably good performance. Its beta may approach that of a standard bipolar device (50 to 100), but it often drops to much lower values if the emitter diffusion is silicided, as is the case in a clad-moat process. Since this transistor injects current into the substrate, care must be taken to provide adequate substrate contact. The main component of collector resistance consists of the lightly doped P-epi layer interposed between the P+ substrate and the PSD diffusion beneath the substrate contacts. A large substrate contact area is necessary to prevent substrate debiasing. A typical CMOS integrated circuit incorporates enough substrate contact in the scribe street to accommodate 10 to 20mA of substrate current. If higher substrate currents will occur, then unused areas of the die should be filled with substrate contacts.

Although a lateral PNP transistor can theoretically be constructed in an N-well process, the absence of NBL encourages substrate injection, and only a small fraction of the total emitter current reaches the intended collector. These transistors exhibit extremely low apparent gains, rendering them of limited use to the analog circuit designer.

Resistors

The most useful of the four resistors available in polysilicon-gate CMOS consists of doped polysilicon (Figure 3.32). Although gate polysilicon exhibits a sheet resistance of only 20 to 30Ω/□, very narrow widths and spacings allow substantial resistance per unit area. A 2μm process can produce polysilicon resistors as area-efficient

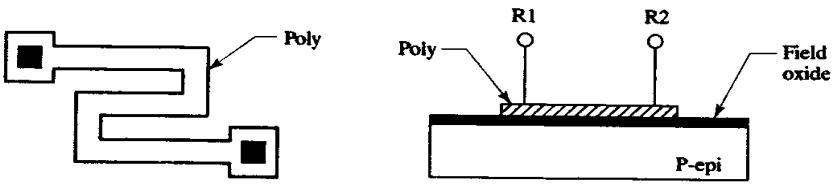


FIGURE 3.32 Layout and cross section of a poly resistor.

as standard bipolar base resistors. Submicron processes can provide remarkable amounts of resistance from narrow polysilicon traces, but the tolerances and matching of such resistors leave much to be desired. In a clad-poly process, a silicide block mask becomes necessary to obtain sufficient sheet resistance to construct practical poly resistors.

The polysilicon resistor of Figure 3.32 consists of a strip of poly deposited on top of field oxide. Contacts at either end allow it to be connected into the circuit. Oxide completely isolates this resistor, enabling it to be biased in any manner desired. Poly resistors can withstand large voltages relative to the substrate (100V or more) and can operate below substrate potential or above the positive supply voltage. The thick-field oxide also reduces parasitic capacitance between the resistor and the underlying substrate. Oxide isolation has one drawback: it isolates the resistor thermally as well as electrically. A poly resistor that dissipates sufficient power will experience permanent resistance variations due to self-induced annealing. Extreme power dissipation will melt or crack polysilicon long before diffused resistors of similar dimensions suffer damage. This behavior allows the construction of polysilicon fuses for wafer-level trimming, but it makes poly resistors undesirable for certain specialized applications, such as ESD protection.

Figure 3.33A shows the layout and cross section of an NSD resistor formed by contacting either end of a strip of NSD diffusion. NSD typically has a sheet resistance of 30 to 50Ω/□. Avalanche breakdown of the relatively shallow NSD diffusion also limits the operating voltage of this resistor, typically to no more than 10 to 15V. A similar resistor can also be constructed from PSD (Figure 3.33B). This resistor consists of a strip of PSD diffusion contained in an N-well region. The well must be biased above the resistor to maintain isolation. The well is therefore connected either to the more positive end of the resistor or to a high-voltage node (for example, the positive supply). PSD resistors also suffer from limited sheet resistance and a relatively low breakdown voltage.

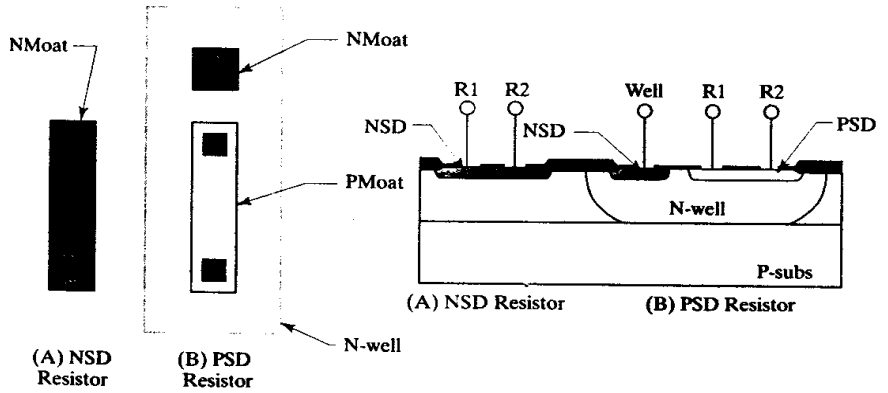


FIGURE 3.33 Layout and cross section of an NSD resistor (A) and a PSD resistor (B).

Another type of resistor consists of a strip of N-well with contacts at either end. NSD placed beneath these contacts prevents the formation of rectifying Schottky barriers. The large spacings required to allow for outdiffusion partially offset the high sheet resistance of the well (typically some 2 to 5kΩ/□). Well resistors are notoriously variable. Slight differences in doping and outdiffusion, voltage modulation of the depletion regions, and surface effects can all cause significant variations in well resistance. Proper layout precautions can help minimize these sources of variability, but most designers prefer to employ narrow polysilicon resistors. Table 3.5 summarizes the properties of the four types of resistors available in a typical CMOS process.

TABLE 3.5 Typical resistor device parameters.

Parameter	Poly	PSD	NSD	Well
Sheet resistance	20Ω/□	50Ω/□	50Ω/□	2000Ω/□
Minimum drawn width	2μm	3μm	3μm	5μm
Breakdown voltage	>100V	15V	15V	50V
Variability (5μm width)	±30%	±20%	±20%	±50% (10μm)

Capacitors

The gate oxide used to fabricate MOS transistors can also be employed to construct capacitors. One plate of the capacitor consists of doped polysilicon and the other of a diffusion, typically N-well. Figure 3.34 shows the layout and cross section of one type of MOS capacitor. The capacitance of this device typically ranges from 0.5 to 1.6fF/μm^2 (0.3 to 1.0pF/mil^2), depending on oxide thickness. The tight control of gate oxide thickness characteristic of modern MOS processes results in typical tolerances of ± 20% as long as the well electrode remains at least 1V above the poly electrode. Failure to maintain adequate bias across the capacitor will cause a dramatic drop in capacitance (Section 6.2.2). The main drawbacks of these capacitors consist of excessive bottom-plate parasitic junction capacitance and series resistance, and of nonlinearity effects at certain voltages.

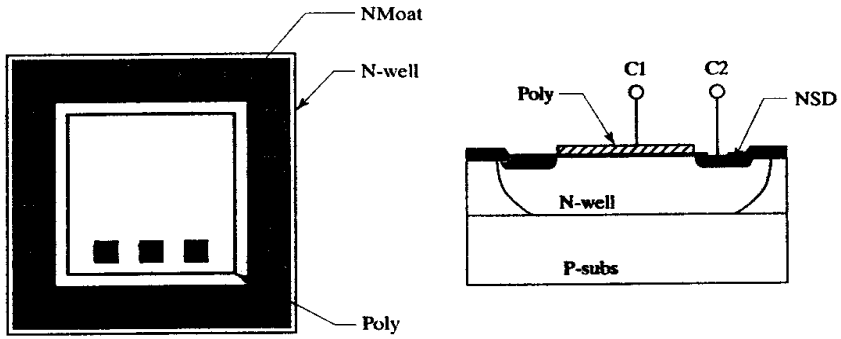


FIGURE 3.34 Layout and cross section of a PMOS capacitor.

3.2.4. Process Extensions

CMOS process extensions tend to focus on improving the PMOS and NMOS transistors rather than on constructing additional devices. One set of extensions seeks to provide higher operating voltages by suppressing hot carrier degradation. Another set of extensions focuses on reducing the size of the transistors to improve packing. Unlike standard bipolar, little emphasis is placed on providing a large number of specialized components because most CMOS fabs build primarily digital products. The large expenditure of time and money required to construct additional devices cannot be justified by the relatively small volume of analog CMOS products. Analog BiCMOS presents an entirely different picture because this process caters specifically to analog design. Many of the features and process extensions of BiCMOS can be retrofitted into a CMOS process if sufficient economic incentive exists.

Double-level Metal

The early CMOS processes used one metal and one poly layer (a combination sometimes called *1½-level metal*). Polysilicon can be used to jumper most signals,

so routing is much easier than for single-level metal. Autorouting software still cannot efficiently route 1½-level metallization, so most modern CMOS processes add a second layer of metal. Analog CMOS layouts can benefit from using this second metal layer to increase packing density. Since planarization becomes more difficult as device sizes shrink. CMOS metallization tends to be substantially thinner than that used for standard bipolar. Higher-power CMOS circuits such as output drivers often benefit from combining multiple metal layers to form a thicker conductor.

Double-level metal adds two mask steps to the process: one for vias and one for metal-2. An *interlevel oxide* (ILO) deposited between the two metal layers provides insulation, and some form of planarization improves planarity for the second-level metal. Although these extra processing steps increase the cost of the wafer, most CMOS fabs routinely employ double-level metal for the majority of their products. Some processes use three, four, or even five metal layers to further reduce the area required for interconnection in high-density autorouted logic.

Silicidation

CMOS processes make extensive use of silicides. In addition to contacts, gate poly is often silicided to reduce its sheet resistance. Some processes even silicide source and drain diffusions to reduce their parasitic resistance. Processes with submicron feature sizes may use all of these forms of silicidation. Older processes with larger feature sizes cannot construct sufficiently fast transistors to justify siliciding their gates and source/drain regions, but they usually employ silicided contacts to prevent contact spiking.

A Schottky diode can be formed on an N-well CMOS process that employs platinum or palladium silicides. The Schottky anode consists of a silicided contact while the cathode consists of N-well contacted by means of an NSD diffusion. The lack of a buried layer increases the internal resistance of this diode and prevents it from being used for high-current applications. The exposed edges of the silicide limit the operating voltage, but techniques exist for suppressing edge breakdown without creating any additional process steps (Section 10.1.3). CMOS processes do not require a Schottky contact mask since a moat region can serve the same function. Note that some silicides do not form usable Schottky barriers. For example, titanium silicide produces such a low forward voltage that the resulting Schottky diodes leak excessively.

Silicidation of the gate poly reduces its sheet resistance to about 2Ω/□, which is too low for constructing most resistors. Poly resistors can still be formed in a silicided-gate process by adding a *silicide block mask*. This mask patterns the silicide layer, either by preventing metal deposition over resistors or by allowing its removal in these areas prior to sintering. The use of a silicide block mask complicates the silicidation process, but is necessary for most analog designs.

Some processes also silicide NSD and PSD diffusions to form so-called *clad moats*. These cannot be used as resistors since their sheet resistances approximate that of their silicide layer (typically about 2Ω/□). A silicide block mask can prevent the silicidation of selected PSD and NSD regions to allow their use as resistors. The silicidation of emitter regions also reduces the beta of substrate PNP transistors (Section 8.3.3). A silicide block mask can sometimes improve substrate PNP beta to the point where this device becomes a practical component.

Lightly Doped Drain (LDD) Transistors

If MOS transistors must operate at high drain-to-source voltages, then precautions become necessary to prevent hot carrier degradation. Typical NMOS transistors

with 3μm channel lengths can indefinitely withstand 5 to 10V, while PMOS transistors of the same dimensions can withstand 15 to 20V. Voltages beyond these limits require alternative structures.

The strong electric field across the pinched-off channel of a saturated MOS transistor causes hot carrier degradation. The electric field intensity diminishes if the depletion region can somehow be widened. In a conventional transistor, the depletion region cannot intrude to any significant extent into the heavily doped drain. If the drain diffusion is more lightly doped, then the depletion region can extend into the drain as well as into the channel and the electric field intensity will decrease. Such *lightly doped drain* (LDD) transistors can withstand substantially higher drain-to-source voltages than can conventional *singly doped drain* (SDD) devices.

LDD transistors actually use two drain diffusions, one forming a lightly doped *drift region* near the edge of the gate, and the other forming a more heavily doped region beneath the contact. The heavily doped diffusion reduces the drain resistance of the structure and allows the transistor to retain most of the performance of a conventional SDD device. One process for forming LDD transistors uses an *oxide sidewall spacer* to self-align the two drain diffusions, enabling precise control of the width of the drift region.[11] Figure 3.35 illustrates the steps required to fabricate this structure. Immediately after patterning the polysilicon gate, a shallow implant self-aligned to the edges of the gate polysilicon deposits the lightly doped drain. The wafer is coated with a thick layer of isotropically deposited oxide. The oxide at the edges of the polysilicon gate is deeper than the oxide over adjacent regions of the wafer. An anisotropic dry etch removes most of the deposited oxide, but some remains along the edges of the gates even after the planar surfaces have entirely cleared. The etch is halted so that filaments of oxide remain along either side of the gate; these form the desired oxide sidewall spacers. These spacers are approximately as wide as the polysilicon gate is thick. A second drain implant self-aligned to the edges of the oxide sidewall spacers forms the heavily doped portions of the LDD structure. The width of the lightly doped drift region approximately equals the width of the sidewall spacer, typically about 0.5μm.

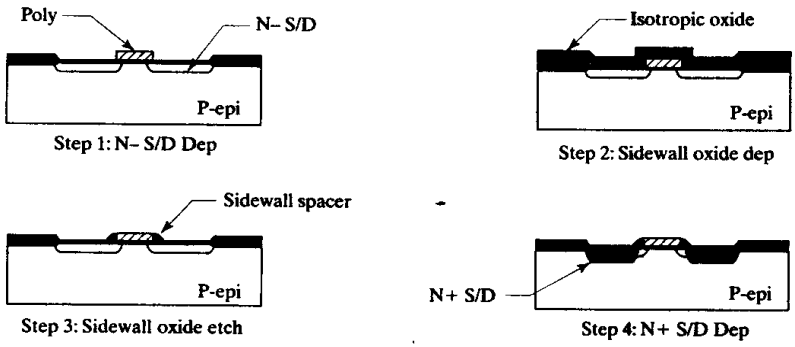


FIGURE 3.35 Steps required to fabricate an LDD NMOS transistor.

Only the drain terminal of the NMOS transistor requires an LDD structure, but no simple way exists to block the formation of the sidewall spacer, so the source terminal of the NMOS also receives an LDD structure. The resulting transistor is *symmetric* in the sense that source and drain can be interchanged without affecting device performance. The PMOS transistor also receives the oxide sidewall spacers.

[11] R. H. Eklund, R. A. Haken, R. H. Havemann, and L. N. Hutter, "BiCMOS Process Technology," in A. R. Alvarez, ed., *BiCMOS Technology and Applications*, 2nd ed. (Boston: Kluwer Academic, 1993), pp. 90–95.

but no lightly doped diffusion. For reasons discussed in section 12.1.1, a channel forms underneath these sidewall spacers. The transistor, therefore, appears to have a slightly longer channel than its drawn dimensions suggest.

The LDD process described previously forms transistors that are suitable for drain-to-source voltages of 10 to 20V. No additional masks are required if all transistors receive both source/drain implants (N– S/D and N+ S/D). Purchasing an additional mask to selectively block the N– S/D implant may provide some additional benefits. Short-channel transistors do not require LDD processing because they break down by punchthrough before hot carrier generation becomes significant. The presence of N– S/D in short-channel NMOS transistors therefore serves no useful purpose. The transistors can pack more densely if the drawn channel dimensions can be reduced without causing premature punchthrough. One way to achieve this goal consists of selectively blocking the N– S/D implant from the short-channel devices. By leaving out the N– S/D, the drawn channel length can be decreased by 0.5 to 1.0μm without affecting the effective dimensions of the device. The purchase of separate N– S/D and N+ S/D masks can make a significant impact on high-density, low-voltage logic circuitry that does not require LDD transistors.

Extended-Drain, High-Voltage Transistors

Oxide sidewall spacers allow the construction of fairly conventional MOS transistors that can withstand drain-to-source voltages of 10 to 20V. Higher voltages require a different approach to constructing the lightly doped drain region to protect against avalanche breakdown as well as hot carriers. Practical high-voltage MOS transistors can be constructed using only the existing masks of the standard N-well polysilicon-gate CMOS process. These *extended-drain* devices do not self-align, so they are inherently long-channel devices that exhibit substantial overlap capacitance. Even so, they suffice for constructing many high-voltage circuits.

Figure 3.36 shows a sample high-voltage, extended-drain NMOS. This transistor uses an N-well region as an extremely lightly doped drain. Since the well is both relatively deep and very lightly doped, it possesses a breakdown voltage in excess of 30V. A plug of NSD provides Ohmic contact to the drain. The source of the transistor consists of an NSD region without the addition of N-well. Since source and drain employ different diffusions, this device is an *asymmetric* MOS transistor.

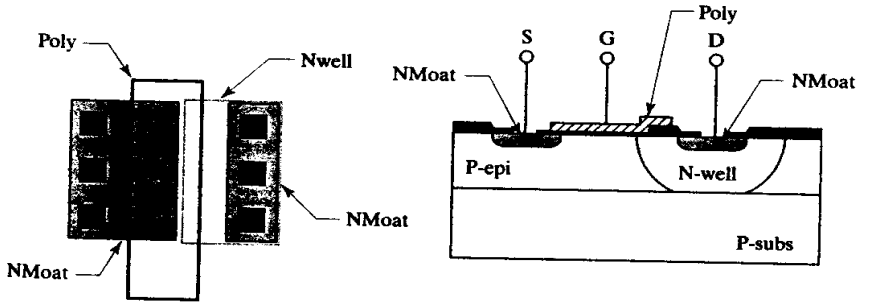


FIGURE 3.36 Layout and cross section of an extended-drain NMOS. The backgate is common to the substrate.

The gate oxide of a high-voltage MOS transistor presents something of a dilemma. The 15V CMOS gate oxide is 300 to 400Å thick and can safely withstand 20 to 25V. Higher voltages will rupture the oxide and destroy the device. A separate gate oxidation can form a thicker oxide for the high-voltage devices, but increasing the oxide thickness decreases the device transconductance. The best solution consists of thickening the gate oxide only over the lightly doped drain where the highest field

stresses occur (Figure 3.36). The thin gate oxide remains in place over the channel and ensures a high device transconductance.

High-voltage, extended-drain PMOS transistors use the P-type channel-stop as a lightly doped drain. These devices are discussed in Section 12.1.2.

3.3 ANALOG BiCMOS

The incessant demand for higher levels of integration has driven the evolution of ever more complex and costly processes. Not only must more devices fit into the same die area, but the performance of these devices must steadily improve to satisfy new applications. By the early 1980s, customers demanded *mixed-signal* integrated circuits incorporating both analog and digital subsystems on a common substrate. A typical mixed-signal integrated circuit contains 90 to 95% digital circuitry and 5 to 10% analog circuitry.[12] CMOS logic overwhelmingly outperforms bipolar logic in packing density and power requirements, so the first attempts to fashion mixed-signal circuits employed unmodified CMOS processes. Analog CMOS circuitry had been designed for decades, so few manufacturers envisioned any difficulty in building the last percentages of the mixed-signal system. But these manufacturers soon discovered that difficulty does not correspond to component count. Although the analog components compose only a few percent of the total devices, they often consume the majority of the design effort. The inferior performance of analog CMOS requires even more design resources to compensate for process inadequacies.

After a few years of failures and qualified successes, most manufacturers began to realize that the analog portions of a mixed-signal system require tailor-made components. The true benefits of using such components lie not in improved performance, but in faster cycle times and higher probabilities of success. The late 1980s saw the development of a new generation of processes specifically aimed at the construction of mixed-signal integrated circuits. These *analog BiCMOS* processes are usually based on CMOS process flows, but are augmented by the addition of bipolar transistors, high-sheet poly resistors, and other special devices.

3.3.1. Essential Features

Analog BiCMOS processes are characterized by their complexity. Most require at least fifteen masks, and the more exotic variants use upward of thirty masks. The penalties of complexity include higher wafer costs, longer manufacturing times, and lower process yields. Set against these disadvantages are the benefits of higher-performance analog circuitry, reduced design effort, and faster design cycles. By the mid-1990s, the majority of new analog designs used some form of analog BiCMOS processing.

Because analog BiCMOS is still an evolving technology, the different manufacturers have not yet settled upon a common definition of the process. Most agree that practical BiCMOS must consist of a standard CMOS flow with the addition of a minimum number of steps to construct adequate bipolar transistors. Most also agree that deep P+ isolation is undesirable because it requires a prolonged high-temperature drive. One alternative form of isolation uses the N-well to form the collector of the NPN transistor. The base and emitter then consist of successive

[12] Reckoned by device count, not area. The relatively few analog devices on a mixed-mode integrated circuit tend to take up an inordinate amount of die area because analog circuitry cannot take full advantage of the reduced dimensions of modern transistors while retaining adequate performance.

counterdopings of the well. The collector-epi junction serves to isolate this style of bipolar transistor, thus the name *collector-diffused-isolation* (CDI).[13] Figure 3.37 shows an NPN transistor constructed using CDI.

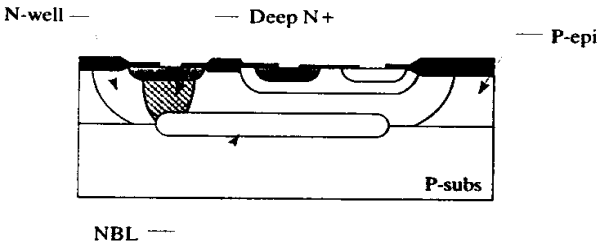


FIGURE 3.37 Details of collector-diffused-isolation (for an NPN transistor). Only the portions of the CDI system related to the collector are labeled.

The specific steps required to construct bipolar transistors on a CMOS process have been hotly debated for some years. While no consensus has yet emerged, analysis shows that three masks are probably needed: NBL, deep-N+, and base.

NBL is the most reluctantly conceded mask, but also the most vital. Some CMOS fabs do not possess epi reactors and lack the experience necessary to implement buried layers. Predictably, attempts have been made to eliminate NBL, and equally predictably, the results have been unsatisfactory. NBL greatly reduces NPN collector resistance—one of the major parasitics of this device. NBL also provides higher NPN operating voltages on the thin epi favored by modern CMOS processing because it blocks vertical punchthrough breakdown. Furthermore, NBL suppresses parasitic substrate PNP action. Without it, lateral PNP transistors become almost impossible to construct. Any practical analog BiCMOS process will almost certainly include NBL.

Standard bipolar uses a deep-N+ sinker to further reduce the collector resistance of power NPN transistors. Although a power MOS transistor can substitute for a power NPN in many applications, deep-N+ remains necessary for creating fully effective minority carrier injection guard rings. Furthermore, due to the graded nature of the well, CDI NPN transistors often exhibit excessive vertical resistance between the collector contact and the NBL. These transistors saturate prematurely, limiting low-voltage operation, complicating device modeling, and causing undesired substrate injection. Most BiCMOS processes include deep-N+, if only as a process extension.

The base diffusion sets the NPN transistor gain, breakdown voltage, and Early voltage. Some processes have attempted to construct NPN transistors using layers scavenged from other devices, with mixed results. Attempts to construct NPN transistors using the diffusions that normally form DMOS transistors have succeeded in some processes but not in others. Questions still remain as to the suitability of the DMOS NPN for applications requiring matching or conducting large currents (Section 12.2.2). Extended-base transistors that use the P-epi or a shallow P-well as a base region have also been successfully constructed, but these devices have several drawbacks. They require more area than conventional CDI devices, especially for small devices, and they often have low Early voltages because they lack a drift region (see Section 8.3.2).

[13] B. T. Murphy, S. M. Neville, and R. A. Pedersen, "Simplified Bipolar Technology and Its Application to Systems," *IEEE J. Solid-State Circuits*, Vol. SC-5, #1, 1970, pp. 7–14.

3.3.2. Fabrication Sequence

This section discusses an analog BiCMOS process based on N-well polysilicon-gate CMOS.[14] The N-well provides collector-diffused isolation; NBL, deep-N+, and base are added to create bipolar transistors. Double-level metal has been added to simplify interconnection. This process requires fifteen masks, one of which is used twice for a total of sixteen masking operations.

Starting Material

The substrate material chosen for analog BiCMOS consists of a P+ (100) substrate cut several degrees off the crystal axis to minimize pattern distortion. The use of NBL in conjunction with a P+ substrate dictates the insertion of an additional epitaxial deposition into the process. Without this step, the NBL would directly contact the substrate to form an N+/P+ junction with a very low breakdown voltage. A lightly doped P-epi layer some 20μm thick therefore resides between substrate and NBL. Three factors determine the thickness of this first epi layer: the up-diffusion of the underlying substrate, the down-diffusion of the NBL, and the width of the depletion region required to withstand the maximum anticipated operating voltage (typically 30 to 50V). The first epi deposits on an unpatterned wafer, so epi-coated wafers may be stockpiled for use as starting material. One could alternatively sacrifice the P+ substrate to eliminate the need for a first epi deposition, but the use of a lightly doped substrate makes the process very susceptible to latchup and substrate debiasing (Section 4.4.1).

N-buried Layer

A brief thermal oxidation grows a thin layer of oxide across the wafer. This oxide is patterned using the N-buried layer (NBL) mask, and an oxide etch opens windows to the silicon surface. Ion implantation deposits an N-type dopant, either arsenic or antimony, in these windows. A brief drive conducted in an oxidizing ambient anneals out lattice damage and causes the formation of the surface discontinuity required for subsequent mask alignment.

Epitaxial Growth

After the NBL anneal, the oxide is stripped and the wafers are returned to the epitaxial reactor for deposition of a second P-epi layer. Surface discontinuities propagate upward through the epi at approximately a 45° angle along a [100] axis determined by the tilt of the wafer. After epitaxial growth, the NBL shadow will have shifted laterally a distance approximately equal to the thickness of the second epi, typically about 10μm. Figure 3.38 shows the wafer after the second epi deposition.

During the growth of the second epi, the reactant gases leach some of the NBL dopant from the surface of the wafer and redeposit it elsewhere, a process called *lateral autodoping*.[15] This mechanism can cause the formation of a thin layer of N-type silicon at the interface between the first and second epi layers, shorting adjacent wells. Autodoping can be limited by using antimony as the dopant or by conducting the epitaxy at reduced pressure. In either case, the BiCMOS NBL is liable to be somewhat more lightly doped than that used for standard bipolar.

---

[14] Eklund, *et al.*, pp. 120ff. Also see J. Erdeljac, B. Todd, L. Hutter, K. Wagensohner, and W. Bucksch, "A 2.0 micron BiCMOS Process Including DMOS Transistors for Merged Linear ASIC Analog/Digital/Power Applications," *Proceedings 1992 Applied Power Elect. Conf.*, 1992. pp. 517–522.

[15] M. W. M. Graef, B. J. H. Leunissen, and H. H. C. de Moor, "Antimony, Arsenic, Phosphorus, and Boron Autodoping in Silicon Epitaxy," *J. Electrochem. Soc.*, Vol. 132, #8, 1985, pp. 1942–1954.

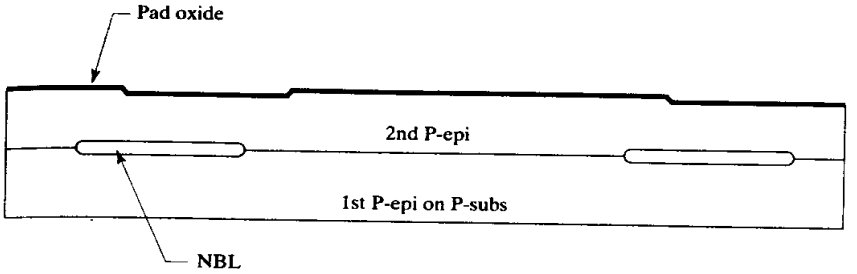


FIGURE 3.38 Wafer after second epitaxial deposition. The P+ substrate is not explicitly shown, but it lies beneath the first P-epi layer.

N-well Diffusion and Deep-N+

A thin layer of oxide is now grown and patterned using the N-well mask. Ion implantation deposits phosphorus, which is subsequently driven down to form the well diffusion. This drive stops before the well and NBL collide to permit the timely insertion of the deep-N+ deposition into the process flow. Additional oxide grown during the well drive allows patterning of the subsequent deep-N+ diffusion. After a heavy dose of phosphorus is implanted, the drive continues until the well and the deep-N+ both overlap the NBL by about 25% of their respective junction depths in order to minimize vertical resistance. Figure 3.39 shows a cross section of the wafer after the drive.

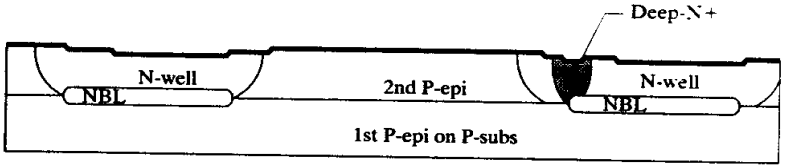


FIGURE 3.39 Wafer after N-well and deep-N+ deposition and drive.

The N-well diffusion influences a number of device parameters for both PMOS and bipolar transistors. Compromises must be made between the two types of devices to the detriment of either or both. For example, short-channel PMOS transistors require a moderately doped well to suppress punchthrough while bipolar transistors need a lightly doped well to form their collector drift region. The doping of the well therefore targets a compromise value that dictates a minimum channel length of 2 to 3μm. If one desires to fabricate shorter channel lengths, then additional wells must be added to allow independent optimization of the MOS and bipolar components of the process.

Base Implant

A uniform, thin layer of pad oxide is grown after the remnants of the previous oxide patterns have been stripped from the wafer. The wafer is patterned using the base mask, and boron is implanted through the pad oxide to form P-type regions, which are subsequently annealed under an inert ambient. Later high-temperature processing steps complete the base drive and set the final base-junction depth. Figure 3.40 shows the wafer after the base anneal. Triple counterdoping degrades the beta of the CDI NPN by raising the total base dopant concentration and hence the recombination rate in the neutral base. This can be partially offset by using a relatively lightly doped base implant. The final base-sheet resistance therefore equals several times that of standard bipolar.

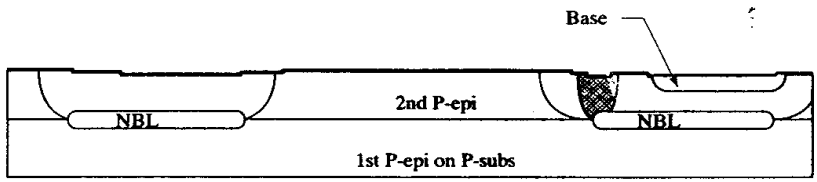


FIGURE 3.40 Wafer after base implant and anneal.

Inverse Moat

Analog BiCMOS uses the same LOCOS technology employed by polysilicon-gate CMOS. A thick layer of LPCVD nitride, patterned using the inverse moat mask, is etched to expose the regions where field oxide will eventually form (Figure 3.41). As in the case of polysilicon-gate CMOS, the MOS transistors occupy moat regions that are not covered by thick-field oxide. Moat regions also serve two additional purposes:

- Base regions are enclosed by moats to prevent oxidation-enhanced diffusion from overdriving the base.
- Schottky contacts are enclosed by moat to allow their etching to proceed simultaneously with base and emitter contacts.

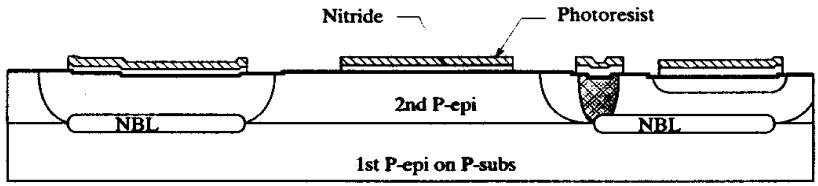


FIGURE 3.41 Wafer after nitride deposition and inverse moat etch.

The inverse moat mask consists of a color reverse of a combination of NMoat, PMoat, base, and Schottky contact drawing layers. Some processes generate the inverse moat mask automatically during pattern generation; other processes require the layout designer to code moat geometries over some or all of the aforementioned drawing layers.

Channel Stop Implants

Since analog BiCMOS uses (100) silicon, channel stop implants are required to raise the thick-field threshold above the operating voltage. The channel stop strategy for analog BiCMOS parallels that for polysilicon-gate CMOS. A blanket boron channel stop adjusts the thick-field threshold over the P-epi, while a patterned phosphorus channel stop sets the thick-field threshold over well regions. The boron channel stop is implanted using the patterned photoresist left from the inverse moat masking operation. A second coat of photoresist is applied and patterned, using the channel stop mask. This mask exposes only the regions of N-well that will ultimately lie beneath the thick-field oxide. A phosphorus implant offsets the previously deposited boron and increases the surface concentration in these well regions. Figure 3.42 illustrates the wafer cross section after the completion of both channel stop implants and the subsequent photoresist removal.

LOCOS Processing and Dummy Gate Oxidation

The LOCOS oxidation employs either steam or elevated pressures to increase the rate of oxide growth. Afterward, the nitride layer and the underlying pad oxide are stripped away. A dummy gate oxidation removes any lingering nitride residue (Figure 3.43).

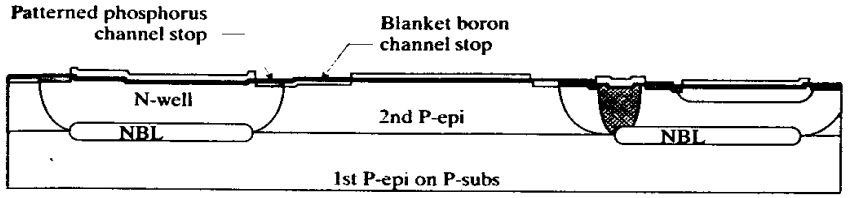


FIGURE 3.42 Wafer after channel stop implants.

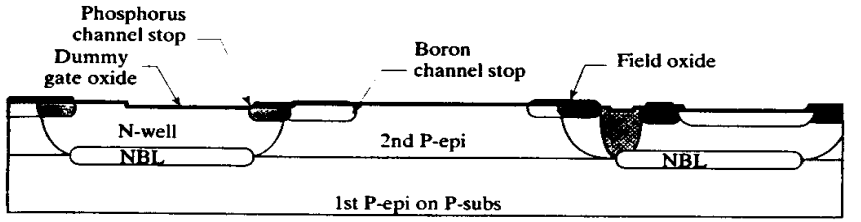


FIGURE 3.43 Wafer after LOCOS oxidation, nitride strip, and dummy gate oxidation.

Threshold Adjust

A single boron V_t adjust implant simultaneously raises the NMOS threshold and lowers the PMOS threshold. With suitable well and epi doping concentrations, both of these thresholds can be simultaneously adjusted to the desired target of 0.7 ± 0.2V. These MOS threshold voltages approximately coincide with the base-emitter voltages of bipolar transistors, although differences in the underlying physics preclude any true matching between the two types of devices.

The threshold implant process consists of coating the wafer with photoresist, patterning the resist with the V_t adjust mask, and implanting the necessary dose of boron through the dummy gate oxide. The final gate dielectric consists of some 300Å of high-quality dry oxide grown after the removal of the dummy gate oxide (Figure 3.43). Figure 3.44 shows a cross section of the resulting wafer.

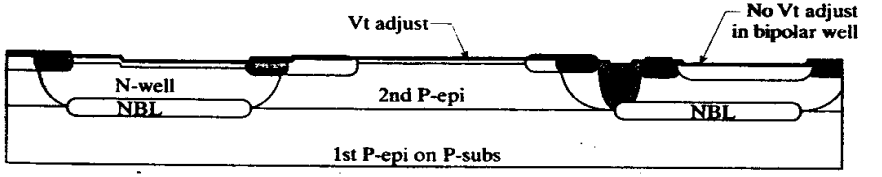


FIGURE 3.44 Wafer after V_t adjust implant.

Polysilicon Deposition and Pattern

The gates of the MOS transistors consist of heavily doped N-type polysilicon formed by depositing intrinsic polysilicon and subsequently doping it with a blanket phosphorus deposition. The patterning step uses the poly-1 mask, so named because a second poly layer may be added as a process option. Figure 3.45 shows the wafer after poly-1 pattern.

Source/Drain Implants

Analog BiCMOS typically produces bipolar transistors with 10 to 20V breakdown voltages. The MOS transistors should ideally be capable of withstanding similar voltages. The breakdown voltages of PSD and NSD can be raised to 15 to 20V without difficulty. Punchthrough can be averted by increasing the minimum channel

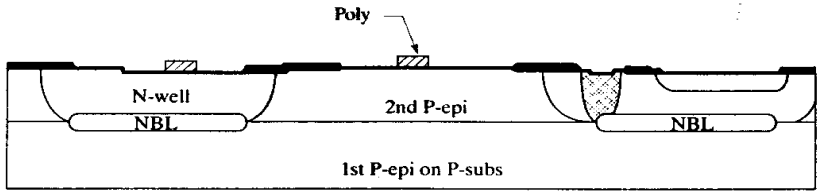


FIGURE 3.45 Wafer after polysilicon deposition and pattern. The channel stop and threshold adjust implants do not appear in this or subsequent cross sections.

length to 2 to 3μm. The addition of a lightly doped drain will suppress hot electron generation in the NMOS transistors.

Photoresist is spun onto the wafer and patterned using the N– S/D mask. A phosphorus implant forms lightly doped source and drain regions that self-align to the polysilicon gate. Isotropic deposition of oxide and subsequent anisotropic etching form sidewall spacers on either side of the gates. A second photoresist layer, patterned using the N+ S/D mask, defines a heavier and somewhat deeper N+ S/D implant that aligns to the edges of the oxide sidewall spacers. The lightly doped drain consists of that portion of the N– S/D implant that lies beneath the sidewall spacers. If all NMOS transistors receive LDD structures, then the N– S/D mask can be reused for the N+ S/D implant.

A 10 to 20V PMOS transistor does not require a lightly doped drain, eliminating the need for a P– S/D implant. The P+ S/D implant occurs after the formation of the sidewall spacers, so the PMOS transistor channel length increases by twice the width of the sidewall spacer (Figure 3.46). This increase in width can be offset by reducing the drawn length of the PMOS gate proportionately.

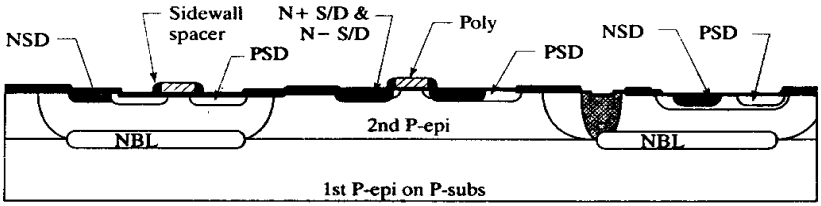


FIGURE 3.46 Wafer after N– S/D, N+ S/D, and P+ S/D implants.

Metallization and Protective Overcoat

The double-level metal flow requires five masks: contact, metal-1, via, metal-2, and protective overcoat. The contacts are silicided to control resistance and, coincidentally, to allow the formation of Schottky diodes. Refractory barrier metal ensures adequate step coverage across the nearly vertical sidewalls of contacts and vias. Copper-doped aluminum minimizes electromigration susceptibility, and a thick layer of compressive nitride protective overcoat provides a mechanical and chemical barrier between the metallization and the encapsulation. Figure 3.47 shows the cross section of a completed wafer, which includes NPN, NMOS, and PMOS transistors.

Process Comparison

Analog BiCMOS uses all of the same steps as polysilicon-gate CMOS. Three additional masks (NBL, deep-N+, and base) are inserted at opportune points in the process as shown by the shaded entries in Table 3.6. NBL deposition must occur before the growth of the second epi. Deep-N+ and base must occur early in the process because these deep diffusions require high temperatures and long drive times.

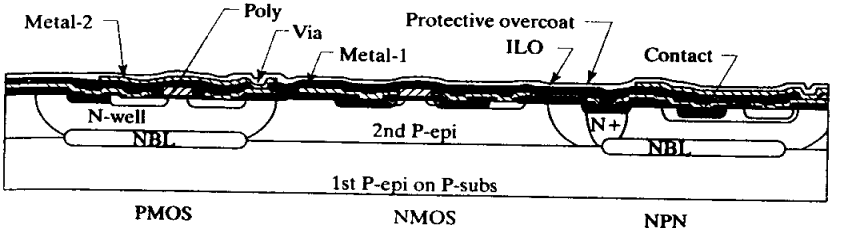


FIGURE 3.47 Completed wafer, showing metal-1 and metal-2 layers.

Mask	Polysilicon-gate CMOS	Analog BiCMOS
1. NBL		First epi growth NBL deposition/anneal
	Epi growth	Second epi growth
2. N-well	N-well deposition/drive	N-well deposition/drive
3. Deep-N+		Deep-N+ deposition/drive
	Pad oxidation	Pad oxidation
4. Base		Base implant/anneal
5. Inverse moat	Nitride deposition/pattern	Nitride deposition/pattern
	Blanket boron channel stop	Blanket boron channel stop
6. Channel stop	Patterned phosphorus channel stop	Patterned phosphorus channel stop
	LOCOS oxidation	LOCOS oxidation
	Nitride/pad oxide strip	Nitride/pad oxide strip
	Dummy gate oxidation	Dummy gate oxidation
7. V_t Adjust	Threshold adjust implant	Threshold adjust implant
	True gate oxidation	True gate oxidation
8. Poly-1	Polysilicon deposition	Polysilicon deposition
	Poly implant/anneal	Poly implant/anneal
9. NSD	N– S/D implant	N– S/D implant
	Sidewall spacer formation	Sidewall spacer formation
10. NSD (again)	N+ S/D implant	N+ S/D implant
11. PSD	P+ S/D implant	P+ S/D implant
	MLO deposition	MLO deposition
12. Contact	Contact OR	Contact OR
	Platinum sputter/sinter/etch	Platinum sputter/sinter/etch
13. Metal-1	1st metal deposition/etch	1st metal deposition/etch
	ILO deposition/planarization	ILO deposition/planarization
14. Via	Via etch	Via etch
15. Metal-2	2nd metal deposition/etch	2nd metal deposition/etch
16. PO	PO deposition/etch	PO deposition/etch

TABLE 3.6 Comparison of analog BiCMOS and polysilicon-gate CMOS processes.

3.3.3. Available Devices

All of the devices available in polysilicon-gate CMOS also exist in analog BiCMOS. These devices include LDD NMOS and SDD PMOS transistors, four types of resistors (poly-1, PSD, NSD, and N-well), gate oxide capacitors, and Schottky diodes. Extended drain transistors can also be created without using any additional masks. Sections 3.2.3 and 3.2.4 discuss the preceding devices in detail. Additional analog

BiCMOS components include a CDI NPN transistor, lateral and substrate PNP transistors, and base resistors.

NPN Transistors

Figure 3.48 shows the layout and cross section of a minimum-geometry NPN transistor. The collector of the NPN consists of an N-well region into which the base and emitter (consisting of NSD) are successively diffused. The inclusion of NBL beneath the active region of the transistor and the addition of a deep-N+ sinker help minimize collector resistance. NSD implanted on top of the sinker ensures Ohmic contact. In a similar manner, PSD allows contact to the lightly doped base. The general appearance of this transistor closely resembles that of the standard bipolar transistor in Figure 3.10. There are, however, several subtle differences.

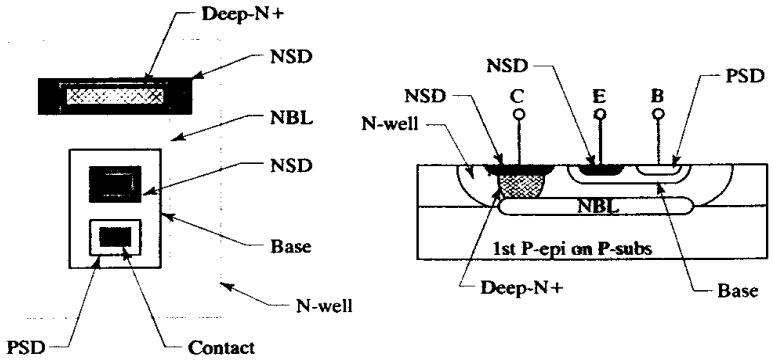


FIGURE 3.48 Layout and cross section of an NPN transistor with deep-N+ and NBL.

The use of three successive counterdopings increases recombination in the neutral base and therefore decreases the gain of the BiCMOS NPN transistor. Lighter base doping partially compensates for this effect. Conventional circuit design techniques require a minimum beta of about fifty. Allowing for a tolerance of ±50%, the nominal beta target becomes seventy-five, which is only about half that offered by a standard bipolar NPN transistor. Higher betas could be achieved, but only at the expense of degrading other device characteristics.

The graded well exhibits an extremely high collector resistance if deep-N+ is not placed beneath the collector contact. This resistance causes a soft transition from the saturation to the forward active region, which resembles the effects of quasisaturation (Section 8.3.2). The transistor may also intrinsically saturate even when the terminal voltages seem to indicate that saturation cannot occur (Section 8.1.4). These problems can be prevented by adding a deep-N+ sinker. Transistors used as Zeners do not require deep-N+, even in analog BiCMOS, because the collectors of these devices do not conduct significant current.

The analog BiCMOS NPN transistor does not perform as well as the standard bipolar NPN transistor, but it still serves for most applications (Table 3.7). Analog BiCMOS also allows much smaller emitter areas than standard bipolar. This benefit does not translate into a proportional reduction in transistor area since many other spacings contribute to the size of the final device. Still, the minimum-geometry analog BiCMOS transistor requires only about 30% of the room of its standard bipolar counterpart.

PNP Transistors

Analog BiCMOS supports both substrate and lateral PNP transistors. Figure 3.49 shows the layout and cross section of a representative substrate PNP transistor. This device is

Parameter	Standard Bipolar	Analog BiCMOS
Drawn emitter area	$100\mu m^2$	$16\mu m^2$
Peak current gain (beta)	150	80
Early voltage	120V	120V
Collector resistance, in saturation	100Ω	200Ω
Typical operating current range for maximum beta	5μA–2mA	1–200μA
V_{EBO} (Emitter-base breakdown, collector open)	7V	8V
V_{CBO} (Collector-base breakdown, emitter open)	60V	50V
V_{CEO} (Collector-emitter breakdown, base open)	45V	25V

TABLE 3.7 NPN device parameters for standard bipolar and analog BiCMOS.

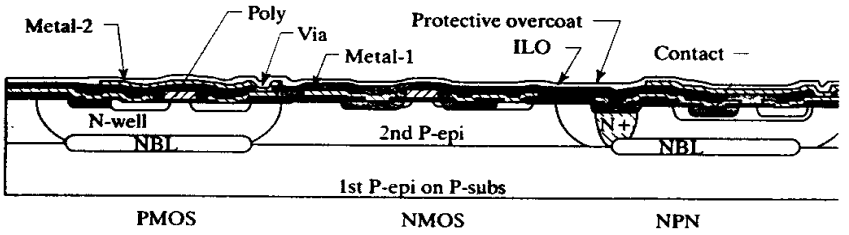


FIGURE 3.49 Layout and cross section of a typical substrate PNP transistor. The substrate acts as the collector.

constructed by implanting PSD into an N-well region. The PSD implant forms the emitter of the transistor while N-well forms the base. A small NSD plug ensures Ohmic contact to the lightly doped N-well. The emitter of the substrate PNP cannot consist of base diffusion because this reaches so deeply into the well that it compromises the breakdown voltage of the transistor. The characteristics of the analog BiCMOS substrate PNP broadly resemble those of a substrate PNP formed in standard bipolar.

Figure 3.50 shows a layout and cross section of a lateral PNP transistor constructed in analog BiCMOS. This device employs the base diffusion to form the emitter and collector, which both reside in an N-well region that forms the base. The addition of NBL to the transistor serves several purposes. First, it acts as a depletion stop, which allows the lateral PNP to withstand higher operating voltages without suffering punchthrough. Second, NBL blocks punchthrough breakdown and allows the base implant to be used in place of the shallower PSD implant. The deeper base implant enhances emitter sidewall injection and therefore raises the beta of the device. NBL also helps to minimize substrate injection. The graded nature of the well reduces the effectiveness of the NBL as a minority carrier barrier, but substantial benefits remain. Without NBL, the lateral transistor would exhibit an apparent beta of less than 10; with NBL the beta can easily exceed 100.

Additional implants must surround the contacts since both the N-well and the base diffusion are too lightly doped to allow direct Ohmic contact. A rectangle of PSD encloses both the emitter and the collector, but this implant only penetrates the moat regions that form the actual collector and emitter regions of the device. The thick LOCOS oxide blocks the PSD implant from reaching the base of the transistor and prevents the P-type implant from shorting the emitter and collector. Similarly, an NSD plug allows Ohmic contact to the N-well. Deep-N+ is not normally

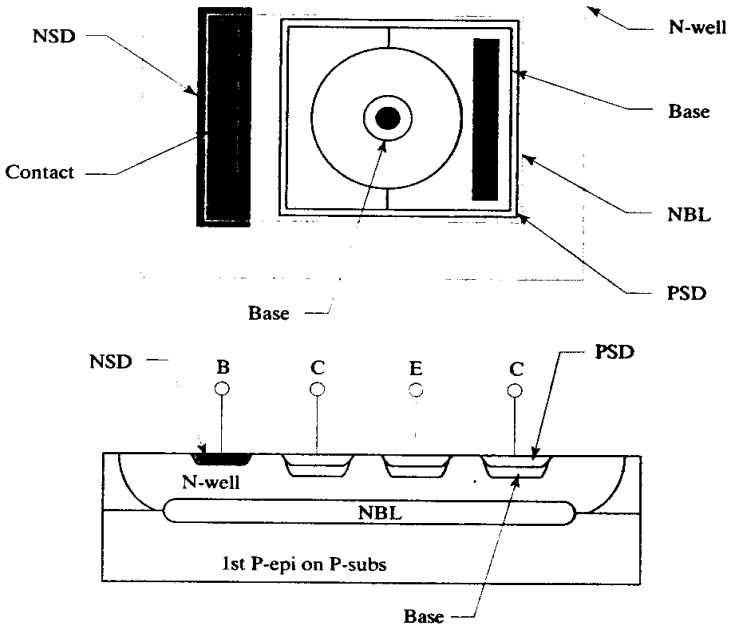


FIGURE 3.50 Layout and cross section of a lateral PNP transistor.

required, although it may become necessary in larger transistors that conduct significant base current.

The minimum-geometry lateral PNP transistor can attain a peak beta well in excess of 100. Fine-line photolithography allows a narrower base width than standard bipolar and greatly reduces the area of a minimum emitter. As the emitter area decreases, the ratio of periphery to area increases. This enhances the desired lateral injection of carriers at the expense of undesired vertical injection. The graded nature of the well and the presence of the channel stop implant generate a doping gradient that forces carriers away from the surface, diminishing surface recombination losses. The use of (100) silicon instead of (111) silicon also reduces surface recombination by minimizing the surface state charge. All of these effects together produce a transistor having a beta of up to 500.

The use of a lightly doped base implant for the emitter reduces emitter injection efficiency and causes a corresponding reduction in beta. Large PNP transistors consist of arrays of many minimum emitters to achieve a high area-to-periphery ratio. Table 3.8 lists several important parameters for both types of PNP transistors available in analog BiCMOS.

TABLE 3.8 Typical PNP device parameters for analog BiCMOS.

Parameter	Lateral PNP	Substrate PNP
Drawn emitter area	16μm²	16μm²
Drawn base width	5μm	—
Peak current gain (beta)	120	100
Early voltage	80V	100V
Typical operating current for maximum beta	1–20μA	1–50μA
V_{EBO} (Emitter-base breakdown, collector open)	45V	45V
V_{CBO} (Collector-base breakdown, emitter open)	45V	45V
V_{CEO} (Collector-emitter breakdown, base open)	30V	45V

Resistors

Analog BiCMOS base resistors consist of rectangles of base material occupying an N-well. Contact is made to either end of the resistor through a PSD plug. The well contact contains an NSD implant to increase the surface doping of the N-well. As with base resistors in standard bipolar, the addition of NBL serves not only to block possible minority carrier injection into the well during transients, but also to raise the operating voltage by preventing punchthrough between the resistor and the underlying substrate. Figure 3.51 shows the cross section and layout of a typical base resistor. The analog BiCMOS base diffusion is relatively lightly doped and typically exhibits a sheet resistance of 500Ω/□. While the high sheet resistance provides compactness, it also complicates resistor layout because the diffusion becomes vulnerable to surface depletion effects (Section 5.3.3). Few designers actually employ base resistors; most prefer to pay the small additional cost for a process extension for high-sheet polysilicon resistors. Pinch resistors are also available, but they offer little or no advantage over minimum-width poly resistors.

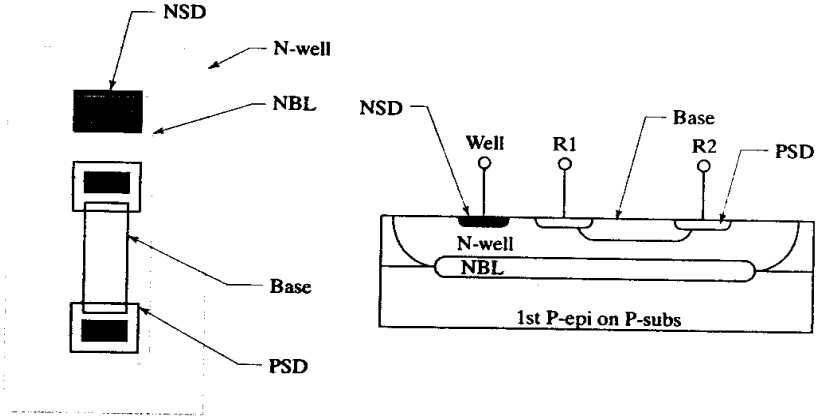


FIGURE 3.51 Layout and cross section of a base resistor.

3.4 SUMMARY

This chapter has examined three representative processes in detail: the standard bipolar process used to construct early analog integrated circuits, a polysilicon-gate CMOS process favored for digital logic, and an analog BiCMOS process, which merges the best of these two technologies onto a common substrate. Variations of these three processes fabricate the bulk of the low-cost, high-volume analog integrated circuits available today.

Standard bipolar is the oldest of the many available bipolar processes, most of which offer much faster switching speeds. The newer processes often replace junction isolation with partial or complete dielectric isolation to simultaneously reduce component area and minimize parasitic capacitance. Diffusions consist of much thinner layers, often created through innovative use of polysilicon as a doping source. Self-alignment and improved dimensional control allow the construction of far smaller geometries. These improvements have increased switching speeds by two orders of magnitude, allowing modern bipolar transistors to operate at speeds approaching fifty gigahertz.[16] The highly specialized processes used to obtain such performance are fully as complex and costly as the most sophisticated of CMOS processes.

[16] These speeds are only obtained using specialized processing techniques, such as the formation of silicon-germanium layers within the base of the transistors.

CMOS processes have undergone their share of evolutionary advances. Gate lengths have shrunk relentlessly as digital designers have sought to pack ever larger numbers of transistors onto limited amounts of silicon real estate. Once gate lengths fall below 1μm, a variety of short-channel effects dictate the use of much more elaborate structures. The backgate doping levels must increase to combat premature punchthrough and elaborate implantation strategies become necessary to confine the channel to the desired region immediately beneath the gate oxide. Cost and complexity have soared as processes push steadily deeper into the submicron regime seeking the elusive ultimate limits of device performance.

BiCMOS technologies seek to merge the best of both worlds. They are therefore heir to both the complexities of submicron CMOS and the elaborate bipolar structures required to obtain high switching speeds. Almost any bipolar or CMOS innovation can be merged into a BiCMOS process, and most have been.

3.5 EXERCISES

Refer to Appendix C for layout rules and process specifications.

3.1. Lay out the NPN transistor shown in Figure 3.10. Use a minimum-size square emitter and a minimum contact overlap of the deep-N+ sinker. Allow room for all necessary metallization.

3.2. Draw a cross section of the NPN transistor shown in Exercise 3.1 to scale, assuming an epi thickness of 10μm, NBL up-diffusion from the epi-substrate junction of 3μm. NBL down-diffusion from the epi-substrate junction of 4μm, an isolation junction depth of 12μm, a deep-N+ junction depth of 9μm, a base junction depth of 2μm, and an emitter junction depth of 1μm. Assume 80% outdiffusion where necessary. Oxide nonplanarities and deposited layers need not be shown.

3.3. Lay out the lateral PNP transistor shown in Figure 3.12. Use the minimum possible basewidth. Allow room for all necessary metallization, including a circular metal field plate connecting to the emitter contact and overhanging the inner edge of the collector region by 2μm.

3.4. Lay out a 500Ω base resistor following the example in Figure 3.13. Make the base resistor 8μm wide, and widen the contacts as much as the width of the resistor allows.

3.5. Lay out a 25kΩ base pinch resistor following the example in Figure 3.15. Assume that all of the resistance comes from the portion of the base beneath the emitter plate. Make the base strip 8μm wide and overlap the emitter plate over the base strip by at least 6μm. NBL should overlap the base strip in the pinched region by at least 2μm.

3.6. Lay out a fingered junction capacitor similar to the one shown in Figure 3.16. Make each of the three emitter fingers 50μm long. The emitter plate should overlap the base by at least 6μm; minimize all other dimensions.

3.7. Lay out the Schottky diode shown in Figure 3.18, assuming a contact opening that is 25 by 25μm. Overlap metal over the Schottky contact layer (SCONT) by no less than 4μm.

3.8. Lay out a 20kΩ HSR resistor. Make the width of the HSR resistor 8μm. The contacts should have the same width as the HSR resistor body. Assume that the base heads contribute negligible resistance and compute the value of the resistor based only on the drawn length of the HSR segment between the base heads.

3.9. Lay out an NMOS transistor with a drawn width of 10μm and a drawn length of 4μm following the example in Figure 3.29. Allow room for all necessary metallization.

3.10. Draw a cross section of the NMOS shown in Exercise 3.9 to scale, assuming a well junction depth of 6μm, PSD and NSD junction depths of 1μm, a gate oxide thickness of 350Å (0.035μm), and a polysilicon thickness of 3kÅ. Ignore the V_t adjust and the channel stop implants. Assume 80% outdiffusion where necessary. Assume the silicon surface is planar, and ignore details of the metallization system.

3.11. Lay out a PMOS transistor with a drawn width of 7μm and a drawn length of 15μm following the example in Figure 3.30. Assume NBL is not used. Include all necessary metallization.

3.12. Lay out a PSD resistor with a value of 200Ω following the example in Figure 3.33B. Make the resistor minimum width, and abut the well contact with one end of the resistor to save space.

3.13. Lay out a 3pF poly capacitor following the example in Figure 3.34. Include all necessary metallization. Assume that contacts and vias can reside on top of the poly plate.

3.14. Lay out the BiCMOS NPN transistor shown in Figure 3.48. The NBL should overlap the base region by at least 2μm. Use minimum emitter dimensions and include all necessary metallization.

3.15. Draw a cross section of the NPN from Exercise 3.14 to proper scale, assuming the dimensions given in Exercise 3.10. Assume NBL diffuses upward by 3μm and downward by 2μm. In addition, assume a first epi thickness of 7μm, a deep-N+ junction depth of 5μm, and a base junction depth of 1.5μm. Ignore channel-stop implants and the effects of LOCOS field oxidation on surface planarity. Assume that the silicon surface is planar, and ignore the details of the metallization system.

3.16. Lay out the BiCMOS lateral PNP transistor shown in Figure 3.50. Assume a minimum basewidth. Unlike the device from Exercise 3.3, this transistor does not require a metal field plate over the base region. NBL should overlap the outer edge of the collector by at least 1.0μm. Include all necessary metallization.

3.17. Lay out the resistor-transistor logic NOR gate shown in Figure 3.52A using standard bipolar layout rules. Place Q_1 and Q_2 in the same tank, and use as small a plug of deep-N+ as possible to contact the collector of this tank. Assume the emitters of Q_1 and Q_2 are both minimum-size. Place R_1 in its own tank. Provide at least one substrate contact. Surround this contact with base: this base region may touch, but not extend into adjacent tanks. Label all inputs and outputs.

3.18. Lay out the CMOS NOR gate shown in Figure 3.52B using poly-gate CMOS layout rules. The *W* and *L* values for each transistor are shown on the schematic in the form of a fraction; 5/3 indicates a drawn width of 5μm and a drawn length of 3μm. Place all PMOS transistors in the same well, and connect this well to V_{DD}. Provide at least one substrate contact. Bring all inputs and outputs up to second-level metal, and label them appropriately.

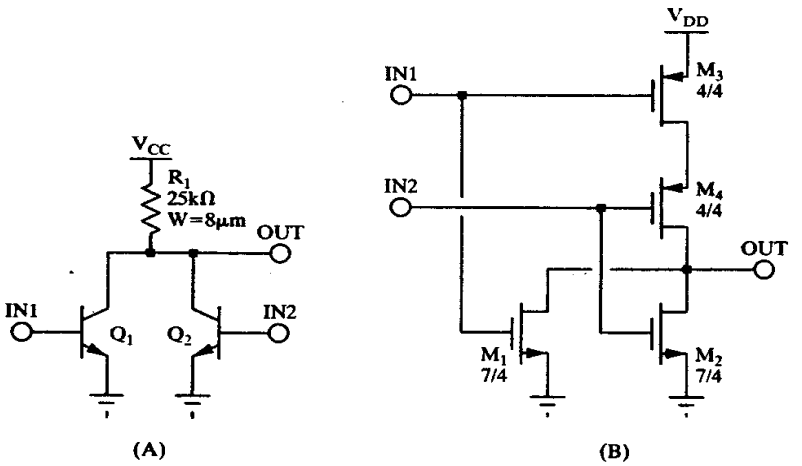


FIGURE 3.52 Circuits for Exercises 3.17 and 3.18.

4 *Failure Mechanisms*

Integrated circuits are incredibly complex devices, and few of them are perfect. Most contain subtle weaknesses and flaws, which predispose them toward eventual failure. Such components can fail catastrophically and without warning after operating perfectly for many years. Engineers have traditionally relied on quality assurance programs to uncover hidden design flaws. Operation under stressful conditions can accelerate many failure mechanisms, but not every design flaw can be found by testing. The designer must therefore find and eliminate as many of these flaws as possible.

The layout of an integrated circuit contributes to many types of failures. If the designer knows about potential weaknesses, then safeguards can be built into the integrated circuit to protect it against failure. This chapter discusses a number of failure mechanisms that can be partially or entirely prevented by layout precautions.

4.1 ELECTRICAL OVERSTRESS

The term *electrical overstress* (EOS) refers to failures caused by the application of excessive voltages or currents to a component. Layout precautions can minimize the probability of three common types of EOS failures. *Electrostatic discharge* (ESD) is a form of electrical overstress caused by static electricity. The addition of special protective structures to vulnerable bondpads can minimize ESD failures. *Electromigration* is a slow wearout mechanism caused by excessive current densities; it can eventually cause open circuits or shorts between adjacent leads. Electromigration failures can be prevented by making leads wide enough to handle the maximum operating currents. The *antenna effect* is an unusual failure mechanism caused by charge accumulation on gate electrodes during etching or ion implantation. The problems posed by the antenna effect can be minimized by following a few simple design guidelines.

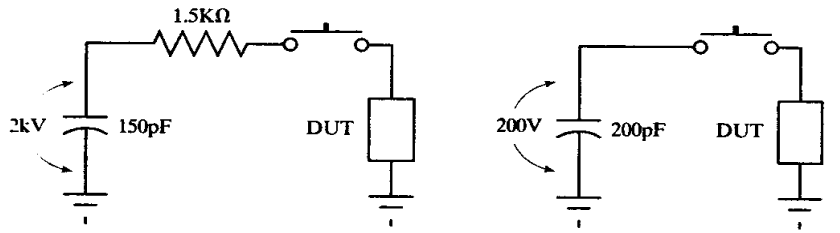
4.1.1. Electrostatic Discharge (ESD)

Almost any form of friction can generate static electricity. For example, if you shuffle across a carpet in dry weather and then touch a metal doorknob, a visible spark will leap from finger to doorknob. The human body acts as a capacitor, and the act

of shuffling across a carpet charges this capacitance to a potential of 10,000V or more. When a finger is brought near the doorknob, the sudden discharge creates a visible spark and a perceptible electrical shock. A discharge of as little as 50V will destroy the gate dielectric of a typical integrated MOS transistor. Voltages this low produce neither visible sparks nor perceptible electrical shock. Almost any human or mechanical activity can produce such low-level electrostatic discharges.

Proper handling precautions will minimize the risks of electrostatic discharge. ESD-sensitive components (including integrated circuits) should always be stored in static-shielded packaging. Grounded wrist straps and soldering irons can reduce potential opportunities for ESD discharges. Humidifiers, ionizers, and antistatic mats can minimize the buildup of static charges around workstations and machinery. These precautions reduce but do not eliminate ESD damage, so manufacturers routinely include special ESD protection structures on-board integrated circuits. These structures are designed to absorb and dissipate moderate levels of ESD energy without damage.

Special tests can measure the vulnerability of an integrated circuit to ESD. The two most common test circuits are called the human body model and the machine model.[1] The *human body model* (HBM) employs the circuit shown in Figure 4.1A. When the switch is pressed, a 150pF capacitor charged to a specified voltage discharges through the integrated circuit to ground. A 1.5kΩ series resistor limits the peak current through the part. Ideally each pair of pins would be independently tested for ESD susceptibility, but most testing regimens only specify a limited number of pin combinations to reduce test time. Each pair of pins is subjected to a series of positive and negative pulses; for example, five positive and five negative. Modern integrated circuits are routinely expected to survive 2kV HBM. Specific pins on certain parts may be required to survive up to 25kV HBM.



(A) Human body model (B) Machine model

FIGURE 4.1 Representative ESD tests: 2kV human body model (A) and 200V machine model (B). In both circuits, DUT stands for *device under test.*

Figure 4.1B shows the circuit employed for the *machine model* (MM). A 200pF capacitor charged to a specified voltage discharges through the integrated circuit to ground. The test circuit contains no intentional series resistance, but practical testers incorporate some 20 to 40Ω in wiring, switches, sockets, and so forth. This, together with the series inductance of the wiring, limits the peak currents generated during testing. The machine model forms a much harsher test than the human body model; few parts can survive more than 500V under these conditions.

A third ESD test called the *charged device model* (CDM) is gradually replacing the machine model. The charged device model charges the integrated circuit package to a specified static potential and then discharges one pin to a low-impedance ground. Researchers believe that this model provides more accurate modeling of

[1] R. Lahri, J. Shibley, D. Merrill, H. Wang, and B. Bastani, "Process Reliability," in A. R. Alvarez, ed., *BiCMOS Technology and Applications*, 2nd ed. (Boston: Kluwer Academic, 1993), p. 170ff.

factory handling conditions than either the human body or the machine model. A typical testing regimen will specify 1 to 1.5kV CDM testing.

Effects

Electrostatic discharge causes several different forms of electrical damage, including gate oxide rupture, gate oxide degradation, and avalanche-induced junction leakage.[2] In extreme cases, ESD discharges can even vaporize metallization or shatter the bulk silicon.

Less than 50V can rupture the gate dielectric of a typical MOS transistor. The rupture occurs in nanoseconds, requires little or no sustained current flow, and is for all intents and purposes irreversible. The rupture typically shorts the gate and the backgate of the damaged transistor.[3] Capacitors that use oxide or nitride dielectrics are also vulnerable to this failure mechanism. An ESD discharge that strikes a pin connecting only to gates or to capacitors may cause oxide rupture in these devices. If the pin connects to any diffusions, then these usually avalanche before the gate oxide ruptures.

The integrity of a gate oxide can be compromised by an ESD event that does not actually rupture it. The weakened oxide can then fail at any time, perhaps after hundreds or thousands of hours of flawless operation. Sometimes the failure does not occur until the product has been delivered to the customer. Testing cannot screen out this type of delayed ESD failure; instead, vulnerable dielectrics must be protected against excessive voltages.

Although junctions are considerably more robust than dielectrics, they can still suffer ESD damage. An avalanching junction dumps a large amount of energy into a small volume of silicon. Extreme current densities can sweep metallization through contacts to short out underlying junctions. Excessive heating can also physically damage junctions by melting or shattering the silicon. These catastrophic forms of junction damage most often manifest themselves as short circuits. Avalanched junctions that do not fail outright usually exhibit increased leakage that may or may not cause the overstressed unit to fail parametric testing.

Preventative Measures

All vulnerable pins must have ESD protection structures connected to their bondpads. Some pins can resist ESD and therefore do not require additional protection. Examples include pins connected to substrate and to large diffusions, such as the collectors of power NPN transistors. These large junctions disperse and absorb the ESD energy before it can damage other circuitry. The power supply pins of most integrated circuits connect to a multitude of diffusions, and thus are also quite robust.

Pins connecting to relatively small diffusions, particularly those connected to the bases or emitters of small NPN transistors, are vulnerable to ESD-induced junction damage. Avalanching the base-emitter junction of an NPN transistor permanently degrades its beta. A circuit designer can sometimes eliminate the vulnerable junctions by rearranging the circuit. ESD protection should be added to any pin connecting to a base-emitter junction, or more generally, to any pin connecting to a relatively small diffusion. Because ESD vulnerabilities are difficult to predict, cautious designers add protection devices to all pins that might be even remotely vulnerable.

---

[2] Some of these are discussed in A. Amerasekera, W. van den Abeelen, L. van Roozendaal, M. Hannemann, and P. Schofield, "ESD Failure Modes: Characteristics, Mechanisms, and Process Influences," *IEEE Trans. Electron Devices*, Vol. 39, #2, 1992, pp. 430–436.

[3] C. M. Osburn and D. W. Ormond, "Dielectric Breakdown in Silicon Dioxide Films on Silicon, II. Influence of Processing and Materials," *J. Electrochem. Soc.*, Vol. 119, #5, 1972, pp. 597–603.

Pins that connect only to gates of MOS transistors or to deposited capacitor electrodes are extremely vulnerable to ESD-induced dielectric rupture. Special gate protection structures should be placed on all such pins. These structures usually include a significant amount of series resistance (typically 500Ω to 5kΩ) and therefore cannot be used on pins that must conduct more than a fraction of a milliamp.

Processes that employ thin emitter oxides are also susceptible to ESD-induced rupture. This vulnerability can be eliminated by ensuring that leads that connect to external bondpads do not cross any emitter region to which they do not connect. Alternatively, ESD structures similar to those used for protecting gates can protect the vulnerable bondpads. Most modern versions of the standard bipolar process employ thick emitter oxides, which eliminate the need for these precautions.[4]

Considerable ingenuity is often required to formulate successful ESD structures for analog integrated circuits. A dozen or more protection circuits are often required to satisfy the large range of voltages and the several types of vulnerable devices found in analog circuits. Section 13.4.3 discusses several commonly employed ESD structures and shows how these can be modified to meet a variety of special requirements.

4.1.2. Electromigration

Electromigration is a slow wearout phenomenon caused by extremely high current densities. The impact of moving carriers with stationary metal atoms causes a gradual displacement of the metal. In aluminum, electromigration only becomes a concern when current densities approach $5{\cdot}10^5 A/cm^2$. Although this may seem a tremendous current density, a minimum-width lead in a submicron process can experience electromigration at currents of only a few milliamps.[5]

Effects

Despite its homogenous appearance, aluminum metal is a polycrystalline material. The individual crystals, or *grains*, normally abut one another. Electromigration causes metal atoms to gradually move away from the grain boundaries, forming voids between adjacent grains. This causes a decrease in the lead's effective cross-sectional area and raises the current density seen by the remainder of the lead. Additional voids form and gradually coalesce until they ultimately sever the lead.

The addition of refractory barrier metal changes the observed modes of failure. Since the refractory metal is relatively resistive, most of the current initially flows through the aluminum. Once voiding finally severs the aluminum, the underlying refractory metal bridges the gap and continues to conduct current. Refractory metals strongly resist the effects of electromigration, so the lead will not completely fail. Instead, the formation of voids in the aluminum causes the lead's resistance to gradually and somewhat erratically increase. More ominously, aluminum metal displaced by voiding sometimes shorts adjacent leads together. The cross-sectional area of the aluminum portion of a lead therefore determines how much current it can safely conduct, regardless of the presence or absence of refractory barrier metal.

Refractory barrier metal is often used to prevent electromigration failures in contacts and vias because it ensures electrical continuity across steep sidewalls after the thin aluminum metallization at these points succumbs to electromigration-induced voiding. Lateral extrusion does not normally occur in contacts or vias since

---

[4] Even thick emitter oxides can rupture under certain conditions; see "Dielectric Breakdown of Emitter Oxide," *Semiconductor Reliability News*, Vol. IV, #1, 1992, p. 1.

[5] Assuming a lead width of one micron and a thickness of 5000Å, a current of 2.5mA will produce a current density of $5{\cdot}10^5 A/cm^2$.

a contiguous sheet of metal covers the entire structure. Likewise, resistance changes caused by voiding are usually small compared to the inherent resistance of the contact or via structure. Where these resistance changes cannot be tolerated, the designer can insert additional contacts or vias to help reduce the current density.

Preventative Measures

The first line of defense against electromigration consists of process improvements. Aluminum metallization is now routinely doped with 0.5 to 4% copper to enhance electromigration resistance.[6] Copper accumulates at the grain boundaries, where it inhibits voiding by increasing the activation energy required to dislodge metal atoms from the lattice. Copper-doped aluminum exhibits five to ten times the current handling capability of pure aluminum.[7] The electromigration resistance of leads can be further improved by using compressively stressed protective overcoats that confine the metal under pressure and inhibit void formation. Refractory barrier metal can also help prevent electromigration failures in contacts and vias. Most manufacturers do not rely upon refractory metal to protect other oxide steps because of the risk of lateral extrusion. Instead, the leads are designed so that the aluminum portion of the metallization can handle the full current over all portions of the metal pattern except at contact and via openings.

Processing techniques can minimize electromigration, but there remains some maximum current density that cannot be exceeded without risking eventual metallization failure. The design rules for each process thus define a maximum allowed current per unit width. Typical values are 2mA/μm (50mA/mil) for leads that do not cross oxide steps and 1mA/μm (25mA/mil) for those that do. These values depend on the thickness of the metallization and its composition, and on the anticipated operating temperature (Section 14.3.3). Consider a lead that must conduct 50mA following the electromigration limits specified above. If this lead routes across field oxide in order to avoid oxide steps, then it need be only 25μm wide; otherwise its width must increase to 50μm. The lead cannot widen abruptly at the oxide steps because the current only gradually flows out from a narrow lead into a wider one. The wider lead should extend beyond the oxide step in either direction for a distance at least twice its greatest width.

Excessive current can also cause bondwires to overheat and fail. In practice, a typical one-mil gold bondwire that is 50mil (1.25mm) in length can safely conduct about an amp, while a similar aluminum wire can conduct about 750mA. If the anticipated currents exceed these limits, then the design will require larger-diameter bondwires or multiple bondwires placed in parallel (Section 14.3.3).

4.1.3. The Antenna Effect

Dry etching uses intense electrical fields to generate an ionizing plasma. During the etching of the gate poly and the oxide sidewall spacers, electrostatic charges may accumulate on the gate poly. The resulting voltages may become so large that current flows through the gate oxide. Although the amount of energy involved is usually inadequate to rupture the gate oxide, it still degrades its dielectric strength. The amount of degradation is proportional to the total charge that passes through the gate oxide divided by the total gate oxide area (Section 11.1.2). Each poly region collects an electrostatic charge proportional to its own area. A small gate oxide

---

[6] I. Ames, F. M. d'Heurle, and R. E. Horstmann, "Reduction of Electromigration in Aluminum Films by Copper Doping," *IBM J. of Research and Development*, Vol. 14, #4. 1970, pp. 461–463.

[7] S. S. Iyer and C. Y. Ting, "Electromigration Study of Al-Cu/Ti/Al-Cu System," *Proc. International Reliability Physics Symp.*, 1984, pp. 273–278. See also Lahri, *et al.*, p. 166.

region connected to a large poly geometry can suffer disproportionate damage. This mechanism is sometimes called the *antenna effect* because the large poly area acts as an antenna to collect the charge that flows through the vulnerable gate oxide. Gate oxide damage due to the antenna effect has also been observed during ion implantation of the source/drain regions.[8,9]

The magnitude of the antenna effect is proportional to the ratio between exposed conductor area and gate oxide area. During the patterning of the polysilicon, the poly is the exposed conductor. Similarly, during the patterning of the first level of metal, the metal is the exposed conductor. Separate area ratios must be computed for each conductor layer. One also computes separate ratios for PMOS and NMOS gate oxides because the two may not break down at exactly the same voltage. Conductor/gate area ratios of several hundred are usually required to produce significant damage. Most layouts contain few geometries with ratios this large, so antenna-effect damage is usually limited to a few locations on the die. Figure 4.2A shows an example of a layout that can produce a conductor-gate area ratio large enough to trigger this type of failure. The gate lead of NMOS transistor M_1 has been elongated to facilitate connection to transistor M_2. The elongated lead has sufficient area to endanger the small gate oxide region of transistor M_1. This vulnerability can be eliminated by inserting a metal jumper in the poly lead next to transistor M_1 (Figure 4.2B). This jumper drastically reduces the area of the poly geometry connected to M_1's gate oxide, which in turn reduces the conductor/gate area ratio to a safe value.

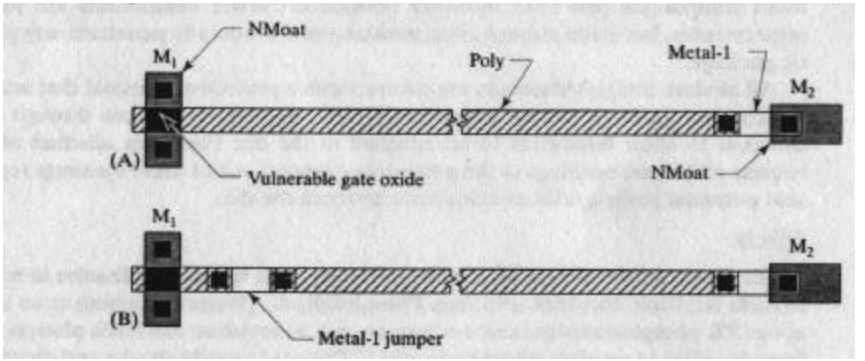


FIGURE 4.2 A layout susceptible to the antenna effect (A) can be made immune by the addition of a metal jumper (B).

Electrostatic damage can also occur during the etching of metal layers. Metal regions connecting to diffusions rarely pose any problem because these diffusions provide paths through which the electrostatic charges can leak away. The topmost layer of metal is almost immune to antenna effects because every geometry on this layer connects to a diffusion somewhere on the die, but the lower metal layers do not necessarily connect to diffusions until the top metal layer is in place. During the etching of a lower metal layer, any geometry that does not connect to a diffusion through layers already present can potentially collect a damaging electrostatic charge.

Antenna effects on lower metal layers can be eliminated by inserting short jumpers on the top metal layer to minimize the conductor area of the lower metal

[8] T. Watanabe and Y. Yoshida, "Dielectric Breakdown of Gate Insulation Due to Reactive Ion Etching," *Solid State Technology*, Vol. 27, #4, 1984, pp. 263–266.

[9] K. Markus, C. M. Osburn, P. Magill, and S. M. Bobbio, "Thin-oxide Degradation Along Feature Edges During Reactive Ion Etching of Polysilicon Gates," *J. Vac. Sci. Technol. A*, Vol. 12, #4, 1994, pp. 1339–1345.

layer attached to small gate oxide regions. This solution resembles that shown in Figure 4.2B. In cases where top-metal jumpers are not feasible, damage can still be avoided by ensuring that the lower-metal lead connects to a diffusion. If the layout does not include any conveniently located diffusion, consider adding a minimum-size NMoat/P-epi or PMoat/N-well diode (Section 10.2). These diodes may affect circuit operation, so they must not be added without consulting the circuit designer.

4.2 CONTAMINATION

Integrated circuits are vulnerable to certain types of contaminants. Assuming that the device has been properly manufactured, very low levels of contaminants will initially exist inside the plastic encapsulation. Plastic mold compounds have been carefully formulated to provide the highest possible degree of resistance to penetration by external contaminants, but no plastic is entirely impregnable. Contaminants seep in along the interface between the metal pins and the plastic, or they directly penetrate the plastic itself. Two major contamination issues faced by modern plastic-encapsulated dice are *dry corrosion* and *mobile ion contamination.*

4.2.1. Dry Corrosion

The aluminum metal system will corrode if exposed to ionic contaminants in the presence of moisture. Only trace amounts of water are necessary to initiate this so-called *dry corrosion.* Since moisture and ionic contaminants are both ubiquitous, integrated circuits must depend on their encapsulation to protect them. Early mold compounds had little moisture resistance. Newer compounds are more impermeable, but given enough time, moisture will eventually penetrate any plastic package.[10]

All modern integrated circuits are covered with a protective overcoat that acts as a secondary moisture barrier. Unfortunately, openings must be made through this overcoat to allow bondwires to be attached to the die. Fuse trim schemes often require additional openings in the protective overcoat. All of these openings represent potential pathways for contaminants to reach the die.

Effects

Water alone cannot corrode aluminum, but many ionic substances dissolve in water to form relatively corrosive solutions. Phosphosilicate glasses containing more than about 5% phosphorus represent a corrosion risk, as moisture can leach phosphorus from the glass to produce phosphoric acid.[11] This acid rapidly attacks and dissolves aluminum, causing open circuit failures. Many modern processes use nitride or oxynitride protective overcoats to ensure that moisture cannot reach the phosphosilicate glass that lies beneath. Alternatively, the phosphorus content of the glass can be reduced by using a combination of boron and phosphorus as dopants. Both of these elements reduce the softening point of a glass, so a *borophosphosilicate glass* (BPSG) will require less phosphorus to achieve the same softening point as a phosphosilicate glass.

Halogen ions in water solution can also corrode aluminum.[12] Common salt, or sodium chloride, provides an abundant source of chloride ions. Moisture seeping

---

[10] J. E. Gunn and S. K. Malik, "Highly Accelerated Temperature and Humidity Stress Technique (HAST)," *Reliability Physics, 19th Annual Proc.*, 1981, pp. 48–51.

[11] W. M. Paulson and R. W. Kirk, "The Effects of Phosphorus-Doped Passivation Glass on the Corrosion of Aluminum," *Proc. Reliability Physics Symposium,* 1974, pp. 172–179.

[12] M. M. Ianuzzi, "Reliability and Failure Mechanisms of Non-hermetic Aluminum SIC's in an Environment Contaminated with Cl_2," *(sic), IEEE Trans. Comp. Hyb. Man. Tech.*, 6, 1983, pp. 191–201.

into the package of an integrated circuit can transport chloride ions to the surface of the die where they can begin to corrode the aluminum metal system. Significant levels of bromides do not normally occur in the environment, but plastic encapsulants often contain organobromine flame retardants. These flame retardants begin to decompose and release bromide ions at temperatures in excess of about 250°C, which limits the storage and soldering temperatures these packages can safely withstand.[13]

Preventative Measures

Although contamination may seem completely beyond the control of the layout designer, several measures can be taken to minimize vulnerabilities in the protective overcoat. The designer should minimize the number and size of all PO openings. A production die should not include any openings that are not absolutely necessary for its manufacture. If the designer wishes to include additional testpads for evaluation, then these should occupy a special test mask. When the part is released to production, the test mask should be replaced by a production PO mask that seals the test pads under protective overcoat.

Metal should overlap bondpad openings on all sides by an amount sufficient to account for misalignment. The metal bondpads will then protect the underlying oxide from the entry of moisture and other contaminants. Openings made for polysilicon or metal fuses should be made as small as possible, and no circuitry of any sort except the fuse element itself should appear within the opening.

4.2.2. Mobile Ion Contamination

Many potential contaminants dissolve in silicon dioxide at elevated temperatures, but most lose their mobility at normal operating temperatures because they become bound into the oxide macromolecule. The alkali metals are exceptions to this rule and remain mobile in silicon dioxide even at room temperature.[14] Of these so-called *mobile ions*, sodium is by far the most common and the most troublesome.

Effects

Mobile ion contamination induces parametric shifts, most noticeably in MOS transistor threshold voltages. Figure 4.3A shows the gate oxide of an NMOS transistor contaminated by sodium during manufacture. The positively charged sodium ions

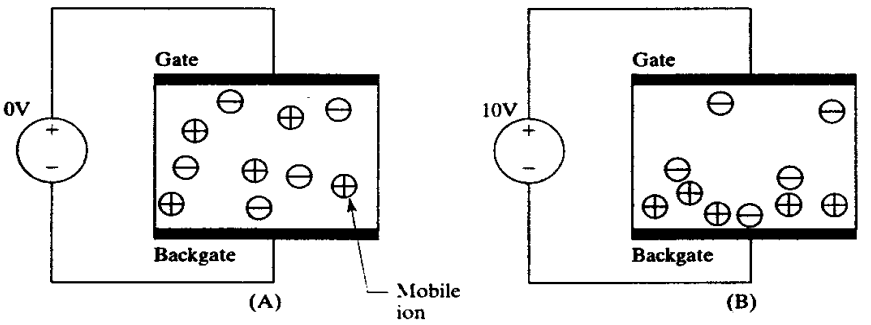


FIGURE 4.3 Behavior of mobile ions under bias: ions that were randomly distributed through the oxide (A) shift in unison in response to a positive gate bias (B).

[13] T. Raymond, "Avoiding Bond Pad Failure Mechanisms in Au-Al Systems," *Semiconductor International*, Sept. 1989, pp. 152–158.

[14] Actually, only lithium, sodium, and potassium (and perhaps hydrogen) qualify as mobile ions in silicon. The heavier alkali metals rubidium and cesium are far less mobile: B. E. Deal, "The Current Understanding of Charges in the Thermally Oxidized Silicon Structure," *J. Electrochem. Soc.*, Vol. 121, #6, 1974, pp. 198C–205C.

are initially distributed throughout the oxide. An equal number of negatively charged ions (anions) are also introduced. Unlike the sodium ions, these anions remain rigidly locked into the oxide macromolecule.[15]

Figure 4.3B shows the same gate dielectric after an extended period of operation under a positive gate bias. The positively charged gate electrode has repelled the mobile sodium ions down toward the oxide-silicon interface. Since the negative ions do not move, the redistribution of the sodium ions results in a net separation of charges within the oxide.[16] The presence of positive charges near the channel of the NMOS transistor decreases its threshold voltage. The magnitude of the threshold voltage shift depends on sodium ion concentration, gate bias, temperature, and time. Many analog circuits require that threshold voltages match within a few millivolts. Even low concentrations of mobile ions can produce shifts of this magnitude.

Mobile ion contamination can produce long-term failures when the slow drift of threshold voltages eventually causes a circuit to exceed its parametric limits. If the faulty devices are removed from operation and are baked at 200°C for a few hours, the mobile ions redistribute and the threshold shifts vanish. This treatment is only temporary; the threshold drift returns as soon as electrical bias is restored. Although analog circuits are particularly susceptible to parametric shifts caused by mobile ions, even digital circuits will eventually fail if the threshold voltages shift too far. Early metal-gate CMOS logic was plagued by threshold voltage shifts caused by severe sodium contamination.

Preventative Measures

Some mobile ions inevitably become incorporated in an integrated circuit during manufacture. This source of contamination can be minimized by using purer chemicals and improved processing techniques. MOS processes typically take extraordinary steps to ensure process cleanliness, but these alone cannot entirely eliminate threshold voltage variations.

Manufacturers of metal gate CMOS attempted to stabilize threshold voltages by adding phosphorus to the gate oxide.[17,18] Phosphorus stabilization had the desired effect of immobilizing alkali metal contaminants, but it also introduced a new problem. The electrically charged phosphate groups shift slightly under strong electrical fields even though they are bound to the oxide macromolecule. Phosphosilicate glasses therefore exhibit the same problem that they were intended to cure! All is not lost, though, because this *dielectric polarization* is not as severe a problem as mobile ion contamination. The threshold shift caused by a given voltage bias remains relatively small—a few tens of millivolts. The threshold shifts caused by dielectric polarization are also much more predictable than those produced by mobile ions, so circuit designers can predict whether a given circuit configuration will be adversely affected or not.[19] The threshold voltage shifts were finally eliminated altogether by using phosphorus-doped polysilicon gates rather than phosphorus-doped gate oxides. Phosphorus-doped polysilicon immobilizes

---

[15] Negative countercharges may not always exist, particularly if the contaminants enter the oxide after manufacture. Threshold shifts still occur because of image effects produced by the presence of electrical charges within the insulating oxide.

[16] N. E. Lycoudes and C. C. Childers, "Semiconductor Instability Failure Mechanisms Review," *IEEE Trans. of Reliability*, Vol. R-29, #3, 1980, pp. 237–249.

[17] M. Kuhn and D. J. Silversmith, "Ionic Contamination and Transport of Mobile Ions in MOS Structures," *J. Electrochem. Soc.*, Vol. 118, 1971, pp. 966–970.

[18] S. R. Hofstein, "Stabilization of MOS Devices," *Solid-State Electronics*, Vol. 10, 1967, pp. 657–670.

[19] E. H. Snow and B. E. Deal, "Polarization Phenomena and Other Properties of Phosphosilicate Glass Films on Silicon," *J. Electrochem. Soc.*, Vol. 113, #3, 1966, pp. 263–269.

alkali metals in much the same way as phosphorus stabilization without the added complication of dielectric polarization.

Moisture seeping into the integrated circuit's package can transport sodium in from the outside environment. Improved packaging materials can slow, but not stop, the ingress of sodium ions. The protective overcoat serves as a further barrier to mobile ions and can prevent them from reaching the vulnerable oxide layers in contact with the silicon. Protective overcoats typically consist of either silicon nitride, which is relatively impermeable to mobile ions, or phosphorus-doped glasses, which can immobilize them. The protective overcoat therefore serves as a final line of defense against impurities entering the die from outside.

Any opening through the protective overcoat represents a potential route for mobile ion contamination to enter the die. The metallization normally seals the bondpad openings, but scars left by probe needles can puncture the metal and expose the interlevel oxide (ILO) beneath. A minimum number of probe pads should be used, and these should be placed around the periphery of the die as far from sensitive analog circuitry as possible. Fuse openings through the protective overcoat also represent vulnerable points that should be kept far away from analog circuitry.

The scribe street surrounding the die typically consists of bare silicon because other materials either fracture or clog the saw blade. Contaminants can seep laterally into exposed oxide layers abutting the scribe street. Special structures, called *scribe seals*, placed around the periphery of the die can slow the ingress of contaminants. Figure 4.4A illustrates a typical scribe seal for a single-level-metal CMOS process. The first component of this scribe seal consists of a narrow contact strip surrounding the active area of the die. This contact must be a continuous ring uninterrupted by any gaps in order for it to block the lateral movement of mobile ions through the field oxide. A P-type diffusion placed underneath this contact allows it to double as a substrate contact. This arrangement is very convenient, as the metal plate forming part of the scribe seal also carries the substrate lead around the periphery of the die. The scribe seal also provides a guaranteed minimum area of substrate contacts.

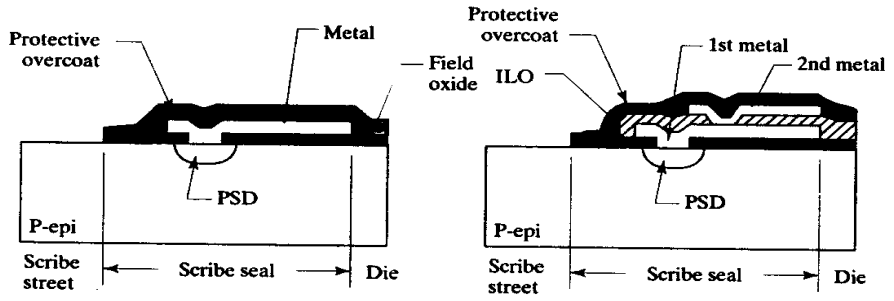


FIGURE 4.4 Scribe seals for single- and double-level-metal variants of a CMOS or BiCMOS process. Depending on the manufacturer, various diffusions may also be placed over the scribe street.

The scribe seal also contains a second contamination barrier formed by flapping the protective overcoat into the scribe street directly on top of the exposed silicon. Any mobile ions attempting to penetrate the scribe seal must first surmount this flap-down and next pass the continuous contact ring before reaching the active regions of the die. Most processes prohibit direct contact between nitride and silicon because compressive stresses in the nitride spawn defects in the silicon lattice. The flap-down of protective overcoat over the scribe street is permitted because the

scribe street does not contain any active circuitry that could be damaged by defects. Nitride should still not touch exposed silicon inside the active area of the die because defects spawned by the damaged silicon can propagate for some distance and may affect adjacent components.

Figure 4.4B shows a scribe seal for a double-level-metal CMOS process. This seal includes a third barrier consisting of a continuous via ring placed just inside the contact ring. This via ring helps prevent contaminants from entering the ILO between the two metal layers. In a triple-level-metal process, a second via ring would be added to protect the second ILO layer between metal-2 and metal-3.

The scribe seals shown in Figure 4.4 can protect almost any die, but the substrate contacts may require different diffusions depending on the process flow. For example, standard bipolar would substitute a combination of P-isolation and base for the PSD rings of Figure 4.4. The P-isolation would probably extend to the edge of the active die because most designers surround the die with a ring of substrate contacts placed underneath the grounded metallization. The functionality of the scribe seals remains the same regardless of the exact diffusions used.

4.3 SURFACE EFFECTS

Surface regions of high electric field intensity can inject hot carriers into the overlying oxide. Surface electric fields can also induce the formation of parasitic channels. Both of these mechanisms are referred to as *surface effects* because they occur at the interface between the silicon and the overlying oxide.

4.3.1. Hot Carrier Injection

Carriers are always in constant motion due to the random thermal vibrations responsible for their diffusion. Electric fields may produce a slow drift of carriers in one direction, but the resulting drift velocity is usually much smaller than the instantaneous velocities produced by thermal agitation. Consequently, electric fields rarely increase the instantaneous velocities of carriers by any perceptible amount (Figure 4.5A). Only at extremely high electric field intensities does the drift of carriers become so rapid that the instantaneous velocities actually increase (Figure 4.5B). The resulting carriers are called *hot carriers* because they move at speeds normally achieved only at elevated temperatures.

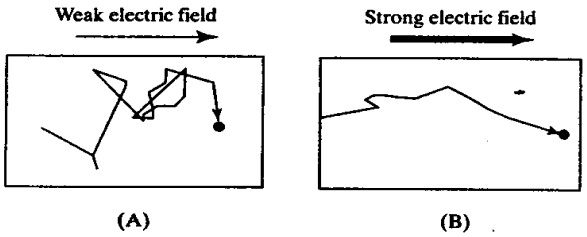


FIGURE 4.5 A weak electric field causes an overall drift of carriers but does not materially affect their instantaneous velocity (A), while a strong electric field actually increases the instantaneous velocity of the carriers (B).

Effects

MOS devices can generate hot carriers when operated in saturation at high drain-to-source voltages. As the drain-to-source voltage increases, the pinched-off portion of the channel slowly grows wider. The increase in width cannot keep pace with the applied voltage, so the electric field intensifies as the voltage increases. At high voltages, the electric field becomes large enough to generate hot carriers near the drain end of the transistor (Figure 4.6). NMOS transistors generate hot electrons, while

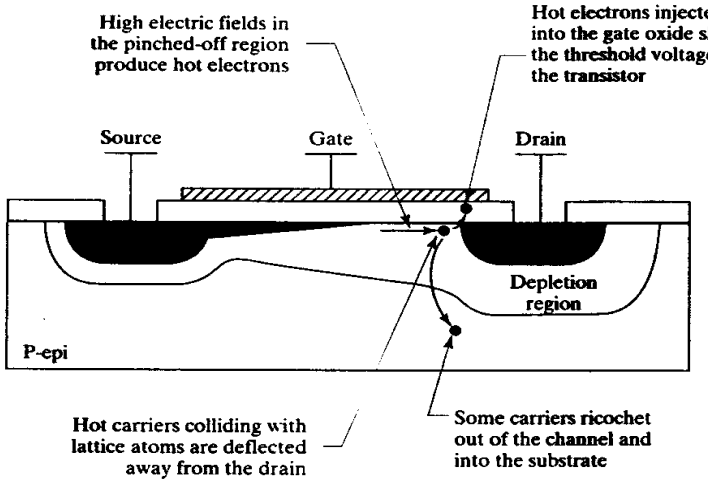

High electric fields in the pinched-off region produce hot electrons
Hot electrons injected into the gate oxide shift the threshold voltage of the transistor
Source
Gate
Drain
Depletion region
P-epi
Hot carriers colliding with lattice atoms are deflected away from the drain
Some carriers ricochet out of the channel and into the substrate

FIGURE 4.6 Simplified diagram showing the mechanism responsible for hot electron injection in an NMOS transistor.

PMOS transistors generate hot holes. Because of differences in effective mobilities, hot carrier production begins at substantially lower voltages in NMOS transistors than in PMOS transistors of similar dimensions. For example, if a 3μm NMOS experiences hot electron injection at 10V, then the equivalent PMOS would probably not experience significant hot hole injection below 20V.

Hot carriers produced at the drain end of the transistor collide with lattice atoms, and some few of the recoiling carriers travel upward into the overlying oxide. Most of these carriers pass through the oxide and return to the silicon, but a few become trapped at defect sites within the oxide. These trapped carriers represent a fixed oxide charge that gradually increases in magnitude as more carriers become trapped. This charge shifts the threshold voltage of the MOS transistor and can in turn affect the performance of the overall circuit.

Parametric shifts caused by hot carriers can be partially or completely reversed by baking the unbiased units at temperatures of 200 to 250°C for several hours. These temperatures impart sufficient thermal energy to the trapped carriers to free them and allow them to return to the silicon. The parametric shifts vanish as the fixed oxide charge dissipates. As in the case of mobile ions, the apparent cure is only temporary. As soon as bias is restored, hot carrier generation resumes and the threshold voltages begin to drift again.

Avalanching junctions also produce large numbers of hot carriers. Avalanche occurs near the surface in most diffused junctions because the dopant concentrations are highest there (Figure 4.7A). Some of the hot carriers produced by the avalanche process travel into the overlying oxide. In the case of a base-emitter junction, these carriers predominantly consist of hot holes that add to the positive fixed oxide charge.[20] As this charge increases, it induces a gradual widening of the depletion region at the surface (Figure 4.7B). The avalanche voltage slowly increases during

[20] G. Blasquez, G. Barbottin, and V. Boisson, "A Review of Passivation-Related Instabilities in Modern Silicon Devices," in G. Barbottin and A. Vapaille, eds., *Instabilities in Silicon Devices, Volume 2: Silicon Passivation and Related Instabilities* (Amsterdam: North-Holland, 1989), pp. 459–460. Blasquez, *et al.* state that Zener walkout in P+/N–junctions spontaneously reverses because some or all of the hot electron charge is not permanently trapped in the oxide macromolecule and can consequently dissipate over time even at room temperature. This effect is usually not observed in N+/P–junctions, presumably because the charge consists largely of holes.

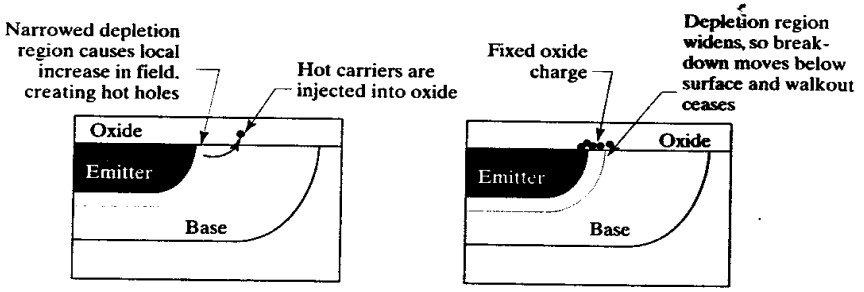


FIGURE 4.7 Simplified diagrams showing Zener walkout mechanism: (A) initial condition of junction, in which hot carrier production occurs near the surface; (B) condition of junction after extended period of operation.

operation, a phenomenon called *Zener walk-out.*[21,22] If a junction diode's reverse breakdown is observed using a curve tracer, the knee of the breakdown curve will gradually "walk out" to higher and higher voltages due to the gradual widening of the depletion region at the surface in response to the accumulation of a fixed oxide charge. Since trapped oxide charges cause walk out, an unbiased high-temperature bake will at least partially reverse it. Depending on processing conditions, emitter-base Zeners can exhibit up to 200mV of walkout.[23] Experimental evidence suggests that the magnitude of Zener walkout diminishes when the process incorporates refractory barrier metal and silicided contacts,[24] although the mechanism responsible for this improvement is not apparent.

Preventative Measures

A *lightly doped drain* (LDD) structure can reduce or even eliminate hot carrier generation in a MOS transistor. Section 3.2.4 discusses the implementation of lightly doped drain structures in a typical polysilicon-gate CMOS process. If no lightly doped drain structure exists, or if the operating voltage exceeds the capabilities of the available structure, then the circuit must be redesigned to reduce the electrical stress imposed on the MOS transistors.

Transistors used as switches generate relatively few hot carriers. Such devices operate either fully on, in which case they are in the linear region, or fully off, in which case they are in cutoff. In neither case does current flow across a large drain-to-source voltage differential. Hot carriers are only generated during brief switching transitions between the two operating states. The average rate of hot carrier generation drops to a minute fraction of the value associated with continuous conduction, and the operating lifetime of the part increases by orders of magnitude. Transistors can withstand voltages far beyond the onset of hot carrier generation as long as switching transitions remain infrequent.

Long channel devices also gain some measure of protection against hot carrier effects. Hot carriers are still produced, but only in the vicinity of the drain. The rest of the channel remains unaffected, minimizing the overall impact of hot carriers on transistor parameters. A few extra volts of operating margin can often be obtained by increasing the channel length a few microns.

---

[21] J. F. Verwey, J. H. Aalberts, and B. J. de Maagt, "Drift of the Breakdown Voltage in Highly Doped Planar Junctions," *Microelectronics and Reliability*, Vol. 12, 1973, pp. 51–56.

[22] R. W. Gurtler, "Avalanche Drift Instability in Planar Passivated p-n Junctions," *IEEE Trans. on Electron Devices*, Vol. ED-15, #12, 1968, pp. 980–986.

[23] W. Bucksch, "Quality and Reliability in Linear Bipolar Design," *TI Technical Journal*, Nov. 1987, pp. 61–69.

[24] W. Bucksch, unpublished manuscript, 1988.

Ordinary base-emitter Zener diodes are surface devices and can therefore exhibit as much as several hundred millivolts of Zener walkout. Attempts have been made to minimize walkout in surface Zeners, but none have been notably successful. The Zener voltage can only be stabilized if the avalanche breakdown is confined to a subsurface region in order to keep hot carriers away from the vulnerable oxide-silicon interface. Such structures are usually called *buried Zeners* (Section 10.1.2).

4.3.2. Parasitic Channels and Charge Spreading

Any conductor placed above the silicon surface can potentially induce a *parasitic channel.* If the conductor bridges two diffusions, then a leakage current can flow through the channel from one diffusion to the other. Most parasitic channels are relatively long and cannot conduct much current, but even small currents can cause parametric shifts in low-power analog circuitry. Channels can sometimes form even in the absence of a conductor due to a mechanism called *charge spreading.* The addition of channel stops or field plates can suppress parasitic channel formation and so protect vulnerable circuitry.

Effects

Both PMOS and NMOS parasitic channels exist. A PMOS parasitic channel can form across any lightly doped N-type region, such as an N-tank in a standard bipolar process or an N-well in a CMOS or BiCMOS process. An NMOS parasitic channel can form across any lightly doped P-type region, such as the P-epi of a CMOS or BiCMOS process, or the lightly doped P-type isolation of standard bipolar processes. Both of these types of parasitic channels can cause a great deal of trouble.

PMOS parasitic channels can form underneath leads crossing lightly doped N-type regions. Consider a metal lead that crosses an N-tank containing a base diffusion (Figure 4.8A). The lead acts as the gate of a PMOS transistor and the N-tank as its backgate. The base region forms the source of the transistor and the isolation serves as its drain. A channel will form if the voltage difference between the lead and the base region exceeds the threshold voltage of the parasitic MOS transistor.[25] Since the thick-field oxide serves as the gate dielectric of this transistor, its threshold voltage is called the *PMOS thick-field threshold.* If the process has a 40V PMOS thick-field threshold, then the base must be biased at least 40V above the lead in order for a channel to form beneath the lead. A similar condition applies to any other potential parasitic PMOS: the P-type region serving as the source must rise

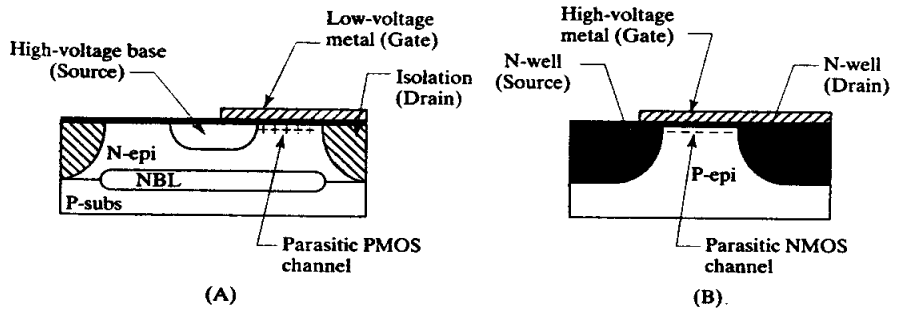


FIGURE 4.8 Parasitic PMOS in a standard bipolar process (A) and parasitic NMOS in an N-well CMOS process (B).

[25] "Bipolar Field Inversion," *Semiconductor Reliability News,* Vol. 3, #1, 1991, p. 7.

above the conductor acting as the gate by a voltage in excess of the PMOS thick-field threshold.

NMOS parasitic channels can form underneath leads crossing lightly doped P-type regions. Figure 4.8B shows a parasitic NMOS channel forming on an N-well CMOS die. This channel forms beneath the lead crossing the lightly doped P-epi. The lead acts as the gate and the P-epi as the backgate. Two adjacent wells serve as the source and the drain. A channel will form if the voltage difference between the gate and the source exceeds the NMOS thick-field threshold. In this case, the voltage on the lead must exceed the voltage on the N-well acting as the source by an amount equal to or greater than the NMOS thick-field threshold. Similar conditions apply to any other potential parasitic NMOS: the conductor serving as the gate must rise above the N-type region serving as the source by a voltage equal to or greater than the NMOS thick-field threshold.

The thick-field threshold voltages of a process depend on a number of factors, including conductor material, oxide thickness, substrate crystal orientation, doping levels, and processing conditions. Most processes quote only one value for the thick-field threshold, this being a minimum value obtained from a worst-case combination of conductors and diffusions. Other processes have undergone more extensive characterization to determine separate thick-field voltages for each combination of conductor and diffusion.

Designers sometimes invoke the body effect (Section 1.4.2) as justification for approaching or even exceeding the thick-field threshold. The body effect increases the apparent threshold voltage of the transistor when the backgate-source junction is reverse-biased. For example, the backgate of the parasitic PMOS in Figure 4.8A is probably biased to a higher voltage than the base. Unfortunately, backgate biasing cannot be relied on for any significant aid. The body effect is most significant in heavily doped backgates, whereas the backgate of a parasitic MOS is usually rather lightly doped. Furthermore, the threshold shift produced by the body effect varies as the square root of the backgate-to-source bias, so even a large backgate bias may not buy more than a few volts of margin.

Engineers once believed that channels could only form beneath conductors, but experience has shown otherwise. Channels can form whenever a suitable source and drain exist, even if no conductor exists to act as a gate. The mechanism underlying the formation of such channels is called *charge spreading*, and although some details still remain unclear, the basic principles are well understood.[26,27] The oxide and nitride films covering an integrated circuit are nearly perfect insulators. Electric current cannot flow through an insulator, but static electrical charges can accumulate on the surface of an insulator or along the interface between two dissimilar insulators. These static charges are not entirely immobile and can slowly shift or spread under the influence of electrical fields. In integrated circuits, the interface between the protective overcoat and the plastic encapsulation is susceptible to this phenomenon. If a nitride protective overcoat is used, then the oxide-nitride interface is also vulnerable. The rate of movement of such charges depends on temperature and on the presence of contaminants. Higher temperatures greatly accelerate charge spreading, as does the presence of even trace amounts of moisture.[28]

Charge spreading requires the presence of static electrical charges at the insulating interface. Experience has shown that these charges do exist and that they consist primarily of electrons, but the mechanisms that generate them are not fully understood.

---

[26] D. G. Edwards, "Testing for MOS IC Failure Modes," *IEEE Trans. Rel.*, R-31, 1982, pp. 9–17.

[27] Lycoudes, *et al.*, p. 240ff.

[28] E. S. Schlegel, G. L. Schnable, R. F. Schwarz, and J. P. Spratt, "Behavior of Surface Ions on Semiconductor Devices," *IEEE Trans. on Electron Devices*, Vol. ED-15, #12, 1968, pp. 973–980.

Hot carrier injection certainly contributes to charge spreading, but integrated circuits that do not produce hot carriers still exhibit charge spreading. Various hypothetical mechanisms have been postulated to account for these experimental observations. In practice, the source of the static charge is less important than its consequences.

Figure 4.9A shows a cross section of a standard bipolar die susceptible to charge spreading. The base region inside the tank is biased above the PMOS thick-field threshold, and therefore acts as the source of a parasitic PMOS transistor. The tank containing this base region is also, of necessity, biased above the PMOS thick-field threshold. Electrons present in the overlying insulating layers will tend to migrate toward the positively charged tank. Eventually, enough electrons may accumulate over the tank to induce a channel (Figure 4.9B). In effect, the static charge generated by charge spreading behaves as the gate electrode of an MOS transistor.

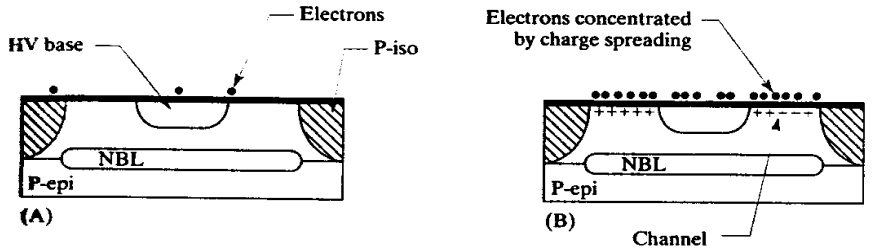


FIGURE 4.9 Cross section of a standard bipolar structure susceptible to charge spreading: (A) before and (B) after an extended period of operation under bias.

Standard bipolar appears to be more susceptible to charge spreading than does CMOS, probably because of less stringent process cleanliness. CMOS processes must minimize ionic contamination to maintain threshold voltage control; no such requirement exists for standard bipolar. The absence of excessive numbers of mobile ions gives CMOS and BiCMOS processes a certain degree of immunity to charge spreading.

Charge spreading produces parasitic PMOS transistors because it involves the accumulation of negative charges. The sources of these parasitic transistors consist of any P-regions that operate at voltages exceeding the PMOS thick-field threshold. The most vulnerable devices contain large, high-voltage P-regions operating at low currents—for example matched high-voltage HSR resistors. Failures tend to occur after long periods of high-temperature operation under bias. Moisture increases the mobility of surface charges, so environmental tests designed to detect moisture sensitivity often uncover charge spreading problems. The resulting parametric shifts resemble those produced by hot carrier injection in that they can be partially or completely reversed by baking the unbiased units at 200 to 250°C for several hours. The high temperature causes the accumulated static charges to disperse and restores an equilibrium between mobile ions and their fixed countercharges. This treatment does not constitute a permanent cure because the parametric drifts resume as soon as bias is restored.

Preventative Measures (Standard Bipolar)

NMOS channel formation can be suppressed in standard bipolar by coding base over all isolation regions. This *base-over-isolation* (BOI) requires no additional die area because the spacings required by the isolation are much larger than those required by the base. The BOI can therefore coincide with the isolation, or even slightly overlap it. Not all standard bipolar processes employ base-over-isolation; some already have a sufficiently heavily doped isolation diffusion to suppress channel formation.

Standard bipolar devices are susceptible to the formation of PMOS channels through charge spreading. Any tank that contains a P-type diffusion biased above the PMOS thick-field threshold requires protection in the form of field plates, channel stops, or a combination of both. Conservative designers usually derate the thick-field threshold of standard bipolar to account for this process's known propensity for charge spreading. For example, a designer might field plate and channel stop high-voltage P-regions operating above 30V even though the process has a rated PMOS thick-field threshold of 40V.

Figure 4.10 shows an example of a high-voltage HSR resistor vulnerable to parasitic channel formation. The tank containing the resistor connects to the positive supply to ensure isolation. A lead must route across the tank to connect to some adjacent low-voltage circuitry. A PMOS channel will form beneath this lead as soon as the voltage difference between the resistor and the lead rises above the PMOS thick-field threshold.

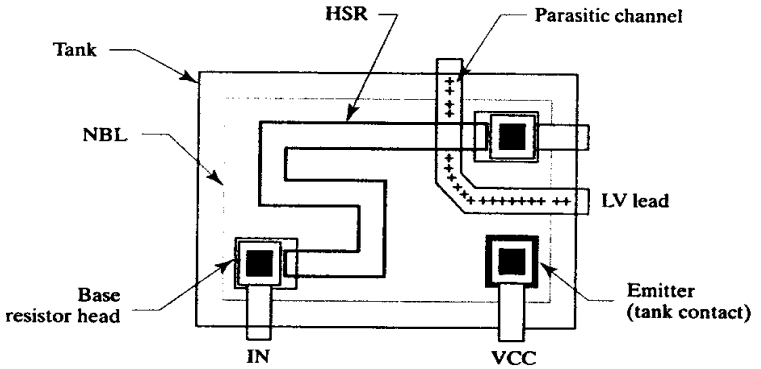


FIGURE 4.10 Example of a circuit susceptible to parasitic PMOS channel formation.

CMOS processes use channel stop implants to raise the thick-field thresholds. Standard bipolar does not include channel stop implants, but an emitter can be coded over selected regions of the N-tank to serve the same purpose. Figure 4.11A shows how emitter diffusions can disrupt the parasitic channels formed beneath a low-voltage lead. Each of the two minimum-width emitter strips disrupts a channel that would otherwise conduct current from the resistor to the isolation.

The emitter bars in Figure 4.11A extend slightly beyond the leads in either direction. These extensions will sever the channel even if the metal and the emitter misalign. The electric field also fringes out to either side of the lead. These fringing

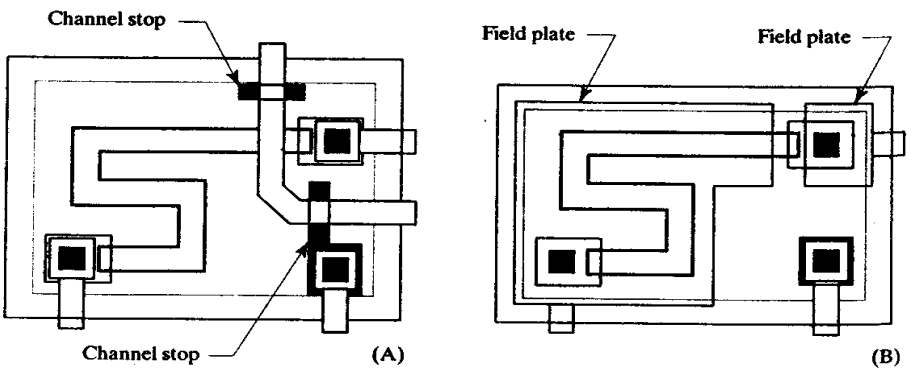


FIGURE 4.11 Two methods for preventing parasitic PMOS channels: (A) channel stops prevent channel formation beneath leads but do not stop charge spreading and (B) field plates provide relatively complete coverage, except possibly in the gap between the plates.

fields rarely extend laterally more than two or three times the oxide thickness, so the overlap of the emitter bar over the lead should equal the maximum photolithographic misalignment plus twice the oxide thickness. Assuming a two-level misalignment of 1μm and a 10kÅ thick-field oxide, the emitter bar should extend about 3 to 5μm beyond the lead on either side. These emitter bars are often called *channel stops*,[29] but they should not be confused with the blanket *channel stop implants* used in CMOS and BiCMOS processes. Channel stops are sometimes called *guard rings*, although this term is more properly applied to minority carrier guard rings (Section 4.4.2).

The channel stops in Figure 4.11A cannot, by themselves, prevent charge spreading. Even if a channel stop entirely encircles a serpentine resistor like that in Figure 4.11, parasitic channels can still form between its turns. If additional channel stops are placed between the turns, parasitic effects could still alter the effective width of the resistor by inverting the silicon along its edges. Some other technique must be used to supplement channel stops, especially for high-voltage diffused resistors.

Field plating can provide comprehensive protection against both parasitic channel formation and charge spreading. A field plate consists of a conductive electrode placed above a vulnerable diffusion and biased to inhibit channel formation.[30] Figure 4.11B shows an HSR resistor with field plates added. The low-voltage lead has been rerouted and a large plate of metal has been placed over the body of the resistor and connected to its positive terminal. The metal lead connecting to the negative end of the resistor has also been enlarged to protect the head of the resistor protruding beyond the main field plate. Both of these field plates must overlap the resistor enough to allow for outdiffusion, misalignment, and fringing fields. Assuming a two-level misalignment of 1μm, a maximum outdiffusion of 2μm, and a maximum fringing distance of 2μm, the total overlap must equal 5μm. Since the field plate consists only of metal, it can extend to fill the required area without enlarging either the resistor or its tank.

A field plate operates by providing an intentional gate for at least a portion of the MOS channel. This gate is biased to prevent the gate-to-source voltage of the parasitic transistor from exceeding the thick-field threshold. The presence of the conductive plate prevents the accumulation of static charges and thus suppresses charge spreading. Field plates also prevent modulation of carrier concentrations in the underlying silicon by acting as electrostatic shields. They therefore provide excellent protection against all types of electrostatic interactions, including conductivity modulation and noise coupling from overlying leads.

Most field plates contain gaps in which channels can still form. In the resistor of Figure 4.11B, a gap remains between the two field plates covering the resistor. Two methods exist for blocking these gaps. One method consists of flaring, or *flanging*, the ends of the field plate to elongate the channel as much as possible (Figure 4.12A). The close proximity of the parallel field plates induces a lateral electric field that sweeps static charges out of this region. The longer the potential channel, the greater the margin of safety provided by the flanges. The second method bridges the gaps between the field plates with short channel stops (Figure 4.12B). The emitter strips used for this purpose must overlap the field plates sufficiently to account for misalignment. This technique combines the strengths of a field plate with those of a channel stop to provide ironclad protection at all points.

[29] J. Trogolo and S. Sutton, "Surface Effects and MOS Parasitics," unpublished report, 1988, p. 13ff.
[30] Trogolo, *et al.*, p. 13ff.

The resistors in Figures 4.11 and 4.12 illustrate another important principle of field plating: the field plate biased to the highest potential should cover as much of the resistor as possible. If the low-voltage field plate were extended, it would provide less protection to the high-voltage end of the resistor. If the voltage difference between the tank and the field plate exceeds the thick-field threshold, then the field plate will actually induce channel formation. The field plate should extend from the high-voltage terminal of a vulnerable resistor to encompass as much of the resistor body as possible. The low-voltage terminal of the resistor should have just enough field plating to cover and protect the contact head. Matched resistors may require a slightly different field-plating strategy (Section 7.2.8).

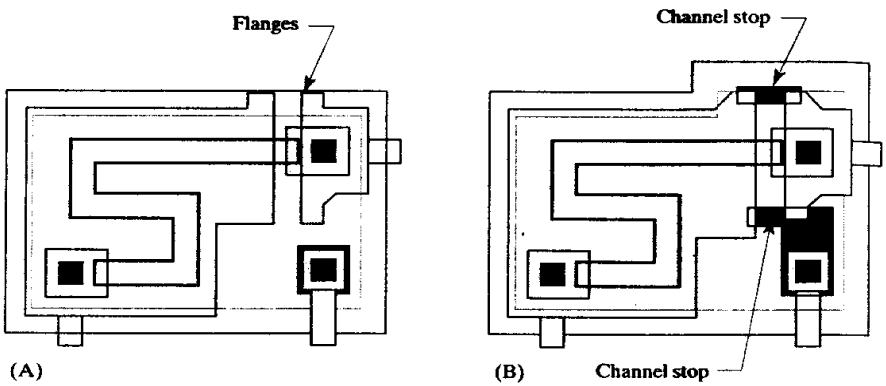


FIGURE 4.12 Improved field plating schemes: (A) flanged field plates and (B) combination of field plates and channel stops.

Figure 4.13 shows an interesting situation that sometimes occurs when laying out resistors. The two terminals of this device connect to a high potential and a low potential, respectively. The high-voltage end of the resistor needs protection against charge spreading, but the low-voltage end does not. Since the voltage drops linearly along the resistor, the field plate has been terminated partway down its length. Partial field plates should extend well beyond the point where the voltage drops below the thick-field threshold. In the case of Figure 4.13, as much of the resistor as possible has been field plated even though much of it apparently serves no useful function. The large safety margin obtained by this means costs nothing and helps ensure that the device will work even under worst-case conditions.

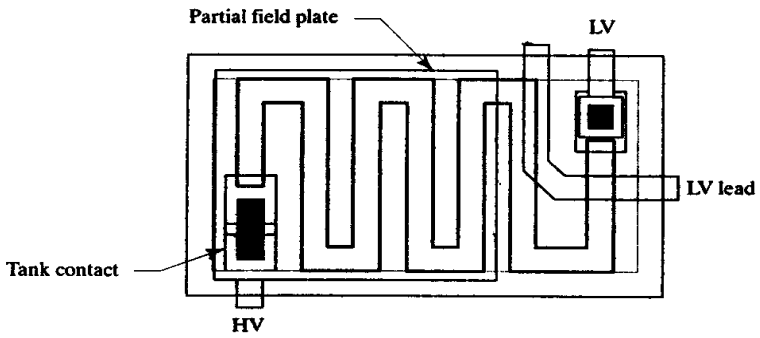


FIGURE 4.13 Example of partial field plating.

Figure 4.14 shows another example of selective field plating involving a multiple-collector lateral PNP transistor. The emitter, base, and one collector operate at voltages in excess of the thick-field threshold, while the remaining collector operates at

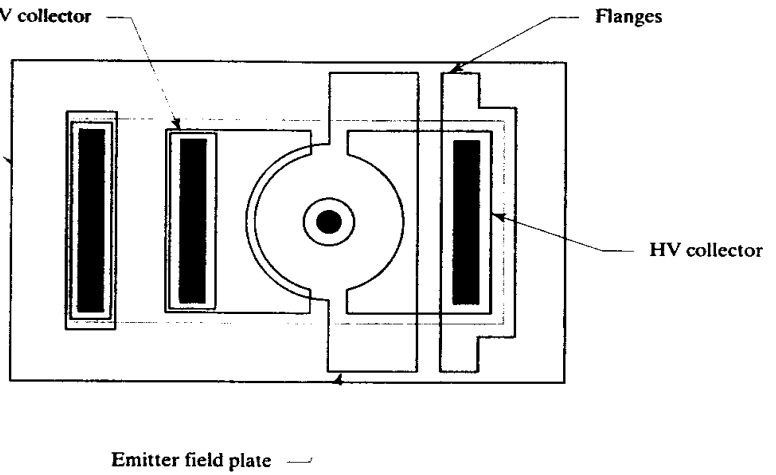


FIGURE 4.14 Field-plated, split-collector, lateral PNP with one low-voltage collector and one high-voltage collector. Flanging suppresses parasitic formation in the gaps between field plates.

a low voltage. The emitter field plate extends out from the emitter across the exposed surface of the base to a point just beyond the inner edge of the collector. The field plate need only overlap the collector by an amount equal to the maximum misalignment minus outdiffusion; certainly no more than 2 to 3μm. A second field plate extends outward from the high-voltage collector to block any parasitic channel that might form between the collectors. or from collector to isolation. No field plate surrounds the low-voltage collector since it does not require one. The field plates have been flanged to ensure that channels cannot form in the gaps. Channel stops could be added, but these would increase the size of the tank and are probably unnecessary.

To summarize, any P-type region biased in excess of the thick-field threshold acts as the source of a parasitic PMOS transistor. Field plates and channel stops ensure that no parasitic channels form from a high-voltage P-type region to any adjacent P-type diffusion. Field plating protects most of the device, while channel stops or flanges protect gaps left in the field plating. The official thick-field threshold of standard bipolar processes should be derated by 25% to provide an additional margin of safety against charge spreading. The following chapters describe additional examples of field plates and channel stops where appropriate.

Preventative Measures (CMOS and BiCMOS)

CMOS and BiCMOS processes usually incorporate channel stop implants to raise the thick-field threshold above the nominal operating voltage. The voltage rating of an N-well CMOS process is usually defined by NSD/P-epi breakdown, PSD/N-well breakdown, or gate oxide rupture, but some structures can withstand much higher voltages. The high N-well/P-substrate breakdown allows PMOS transistors to operate at elevated backgate voltages. These transistors will function normally as long as the drain-to-source voltage does not exceed the PSD/N-well breakdown voltage or the N-well punchthrough voltage. Similarly. an extended-drain NMOS using N-well as a lightly doped drain can withstand the full N-well/P-substrate breakdown voltage applied to its drain terminal.

The lightly doped N-well inverts in much the same manner as the lightly doped N-epi tanks of standard bipolar. Any N-well region containing a P-type diffusion biased above the thick-field threshold becomes vulnerable. As before, PMOS parasitic channels can be suppressed by field plates, channel stops, or a combination

of both. The flanged field plate (Figure 4.12A) is especially attractive because the tighter CMOS layout rules allow a narrower gap between the flanges. The stronger lateral electric field makes close-spaced flanges particularly effective at preventing the accumulation of static charges. Flanged fieldplates are the method of choice for suppressing parasitic PMOS channels in high-voltage CMOS and BiCMOS structures.

Charge spreading is less prevalent in CMOS processes than bipolar ones, probably because of improved process cleanliness. Many CMOS designers consequently take a rather cavalier approach to field plates and channel stops. Such indiscretion is unwise considering the greatly reduced operating currents characteristic of modern CMOS designs. However, one special case does exist where charge spreading can be safely ignored. Most CMOS processes list a lower thick-field threshold for poly than for metal because the oxide layer beneath the poly consists only of thick-field oxide and MLO, while that beneath the metal contains an added layer of deposited ILO. Static charges can only accumulate at the interface between two dissimilar materials, so the lower thick-field thresholds associated with thinner oxides do not have any significance for charge spreading. Charge spreading becomes significant only at voltages beyond the highest thick-field threshold listed for the process.

Poly leads can induce parasitic channels if they run across an N-well containing a P-diffusion biased above the poly thick-field threshold. Figure 4.15A shows a typical example of a vulnerable structure consisting of the poly gate lead from a high-voltage PMOS transistor extending across the well and into the surrounding isolation. The well forms the backgate of the parasitic PMOS, the poly acts as the gate, the sources are the PSD regions of the PMOS transistor, and the drain is the P-epi isolation. As long as the backgate potential does not exceed the metal thick-field threshold, the channel can be interrupted by stopping the polysilicon lead short of the drawn edge of the well (Figure 4.15B). The channel can form only beneath the polysilicon lead, so a complete channel cannot form if the lead does not bridge the gap between source and drain. Charge spreading is unlikely to occur so long as the voltages involved do not exceed the highest thick-field threshold of the process. The minimum spacing between the poly and the drawn edge of the N-well should equal the photolithographic misalignment allowance, plus an extra 2 to 3μm to account for fringing fields. The outdiffusion of the well does not

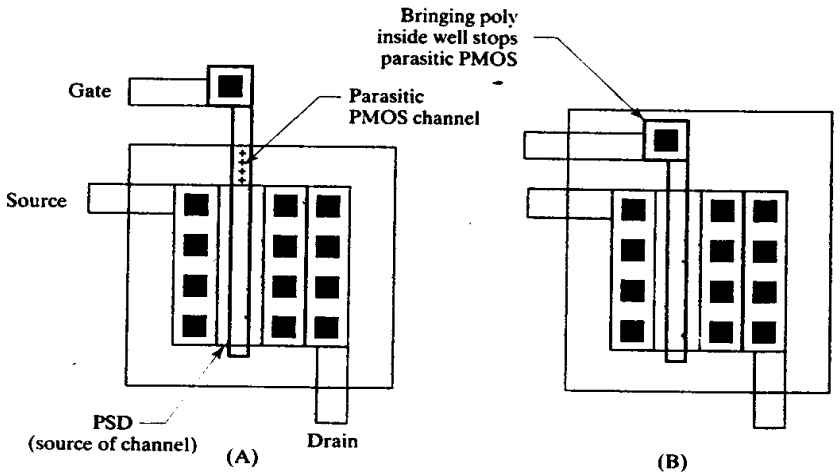


FIGURE 4.15 The parasitic PMOS channel beneath a poly lead (A) can be eliminated by pulling poly inside the well (B).

provide any margin of safety because it becomes very lightly doped beyond its drawn boundaries.

The lightly doped P-type epi can also invert if a high-voltage lead runs across it (Figure 4.8B). The source and drain of this parasitic NMOS consist of two adjacent N-wells; the high-voltage lead acts as the gate, and the P-epi acts as the backgate. A parasitic channel will form if the voltage differential between the lead and an adjacent well exceeds the NMOS thick-field threshold. A channel stop can be inserted by running a thin bar or ring of PSD material down the center of the P-epi beneath the high-voltage lead. The PSD should extend beyond either edge of the lead by an amount sufficient to account for misalignment, plus an additional 2 to 3μm to allow for fringing effects. In many cases, the N-well to N-well spacing can accommodate a minimum-width PSD channel stop with little or no increase in the spacing between adjacent wells. A thin ring of PSD material can then encircle each well (Figure 4.16). This ring not only stops any possible leakage caused by charge spreading but also allows complete freedom to route the leads in any pattern desired. If the PSD rings are drawn when the wells are placed, or if they are automatically produced during mask generation, then the designer can subsequently ignore NMOS channel formation. These PSD rings correspond to the base-over-isolation (BOI) scheme used for the same purpose in standard bipolar designs.

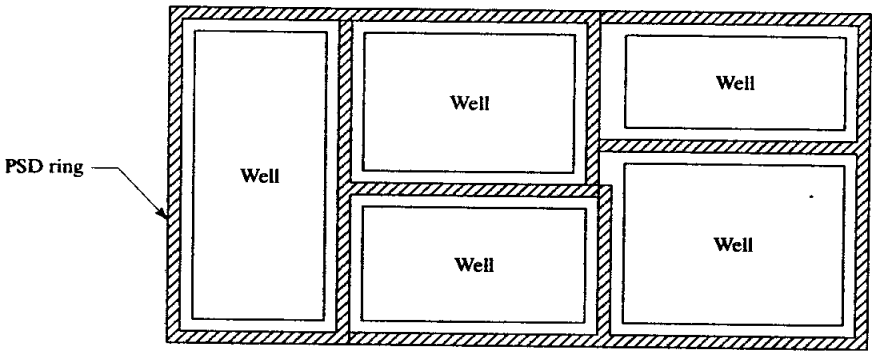


FIGURE 4.16 Sample layout showing the use of PSD rings to prevent NMOS channels.

4.4 PARASITICS

All integrated circuits contain electrical elements not required for their operation. These include reverse-biased isolation junctions, and resistances and capacitances between various diffusions and depositions. The circuit does not benefit from the presence of these *parasitic components*, but they can sometimes adversely affect its operation.

Parasitics are responsible for a number of different types of electrical failures. For example, capacitive coupling can inject noise into sensitive circuitry. The type of parasitics that will be discussed in this section concern the forward biasing of junctions that normally remain reverse-biased. When these junctions forward bias, current begins to flow between circuit nodes that normally remain isolated from one another. If these currents are small and the circuit is relatively insensitive to their presence, then these leakages may produce only subtle parametric shifts. Larger currents can catastrophically disrupt the operation of the circuit. The malfunctioning circuit may actually *latch up*, causing it to continue malfunctioning even after the removal of the triggering event. Latchup can cause physical destruction of an integrated circuit due to excessive power dissipation and consequent overheating. Even if the circuit does not self-destruct, normal operation can only be restored by interrupting the power.

Two important parasitic mechanisms involve currents flowing through the substrate. *Substrate debiasing* occurs when parasitic currents induce voltage drops in a resistive substrate. If these voltage drops become large enough, they can forward-bias one of the isolation junctions. The forward-biased junction then injects current into other circuit nodes, causing potentially catastrophic malfunctions. *Minority carrier injection* occurs when a forward-biased junction injects minority carriers into the isolation, a tank, or a well. Some of these carriers diffuse several hundred microns before recombining and can easily cross reverse-biased junctions that block majority carrier flow.

4.4.1. Substrate Debiasing

Substrate debiasing becomes a problem when currents flowing through the substrate generate voltage drops of a few tenths of a volt or more. This substrate current consists of majority carriers that cannot surmount reverse-biased isolation junctions, but sufficient debiasing may cause one or more isolation junctions to forward-bias and inject minority carriers into active circuitry.

Figure 4.17 shows a typical example of substrate debiasing in a standard bipolar process. Substrate PNP transistor Q_1 injects its collector current I_C directly into the substrate. This current then flows laterally to substrate contact SC_1. Because of the presence of substrate resistance R_s, the substrate voltage immediately under NPN transistor Q_2 rises. Only a few hundred millivolts of substrate debiasing are necessary to forward-bias the collector-substrate junction of a saturated common-emitter NPN.

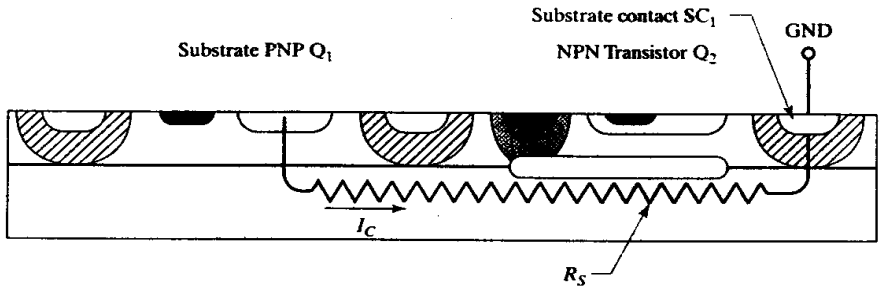


FIGURE 4.17 Cross section of a standard bipolar die showing potential substrate debiasing caused by substrate resistance R_s.

Effects

The voltage required to forward-bias a PN junction depends on both current density and temperature. Table 4.1 lists typical forward-bias voltages for the collector-substrate junction of a minimum-area NPN transistor constructed in a standard bipolar process. This table is useful for estimating susceptibility to substrate debiasing. For example, a circuit using 100μA minimum currents can probably tolerate 1μA of leakage. If it must operate at 125°C, then Table 4.1 indicates that substrate debias-

TABLE 4.1 Forward voltages for a typical collector-substrate junction of a minimum NPN transistor in standard bipolar, as a function of temperature and current.[31]

Current	25°C	85°C	125°C	150°C
10nA	0.43V	0.29V	0.19V	0.13V
100nA	0.49V	0.36V	0.27V	0.22V
1μA	0.55V	0.43V	0.35V	0.30V
10μA	0.61V	0.50V	0.43V	0.39V
100μA	0.67V	0.57V	0.51V	0.47V

[31] Based on V_{BE}(150°C, 1μA) = 0.3V.

ing must not exceed 0.35V. If the same circuit has to operate at 150°C, then it can tolerate no more than 0.30V of debiasing.

Figure 4.18 depicts the cross section of a standard bipolar wafer containing a single substrate current injector and a single substrate contact. R_1 models the lateral resistance through the substrate, while R_2 models the vertical resistance beneath the substrate contact. The total resistance of the substrate R_s equals the sum of the lateral and vertical components: $R_s = R_1 + R_2$.

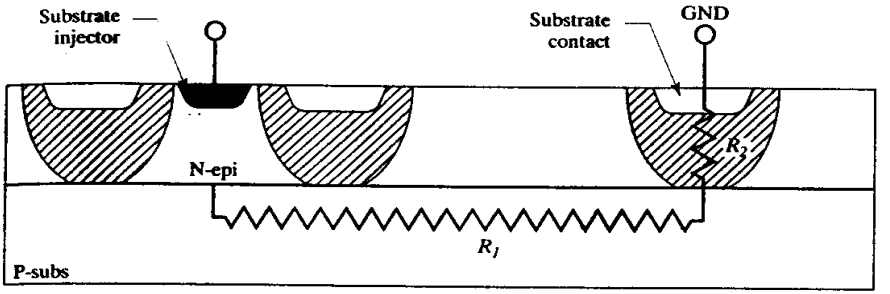


FIGURE 4.18 Simplified model of substrate debiasing in a standard bipolar process.

The relative magnitudes of R_1 and R_2 depend upon the process. Standard bipolar uses a lightly doped substrate and a heavily doped isolation diffusion, so $R_1 >> R_2$. The value of R_1 depends on various geometric factors, including the cross-sectional area of both the injector and the substrate contact as well as the distance between them. Typical values of R_1 range from hundreds to thousands of Ohms,[32] while R_2 rarely exceeds 10Ω.[33] The presence of a network of isolation diffusions criss-crossing the die complicates the computation of substrate resistance because the low sheet resistance of the isolation (usually about 10Ω/□) allows each substrate contact to extract current from a large area of isolation. This effect complicates the computations to such an extent that only empirical measurement or sophisticated computer simulation can yield accurate results.

Two points should be kept in mind when using lightly doped substrates. First, substrate resistance always increases with separation. A substrate contact placed adjacent to the injector will extract some of the current before it ever reaches the substrate. Contacts placed further away require the current either to flow through long stretches of isolation or to pass through the highly resistive substrate. Second, a substrate contact to a heavily doped isolation diffusion draws current not only from the substrate immediately beneath it but also from adjoining stretches of isolation. This effectively magnifies the area of substrate contacts, allowing even a minimum-size contact to have an effective area of many hundreds of square microns. Because of this effect, a scattering of minimum-size substrate contacts throughout the die will have a much lower effective resistance than a single large contact.

---

[32] An estimate of the lateral resistance R_1 can be obtained by examining the spreading resistance $R_{sp} = \rho/d$, where ρ is resistivity and d is the diameter of the points of contact. Spreading resistance assumes a semi-infinite slab of uniformly doped material and a probe separation much wider than the probe diameter: these conditions are only approximately met by typical substrate contacts. Assuming that the cross-sectional areas of injector and substrate contact are $1mil^2$ each, this yields $d = 28.7\mu m$. A substrate with a resistivity of 10Ω·cm would have a spreading resistance of 3.48kΩ. More accurate results can be obtained by applying various correction factors: G. A. Gruber and R. F. Pfeifer, "The Evaluation of Thin Silicon Layers by Spreading Resistance Measurements," *National Bureau of Standards Special Publication 400–10*, Spreading Resistance Symposium, NBS, Gaithersburg, Maryland: June 1974.

[33] The vertical resistance through single-diffused isolation can be approximated by dividing the diffusion into multiple layers of constant doping. Computations for a diffusion with a surface doping of $10^{20}cm^{-3}$, a minimum dopant concentration of $10^{17}cm^{-3}$, and a depth of 5μm yield a resistance of about $4\Omega/mil^2$.

CMOS and BiCMOS processes usually employ a heavily doped substrate and a lightly doped epi, so $R_1 << R_2$. The value of R_1 is usually so small that it can safely be ignored. The value of R_2 depends on the thickness of the epi layer and its resistivity. A typical value is about 600kΩ/μm² (1kΩ/mil²). This value can be used to compute the area of substrate contacts required for a CMOS or BiCMOS design, as explained in the following section.

Even on a heavily doped substrate, contacts placed immediately adjacent to a substrate injector will exhibit less resistance than ones placed far away. This *proximity effect* falls off rapidly with distance, and substrate contacts placed hundreds of microns away are no more effective than those placed on the opposite side of the die. The proximity effect occurs because carriers can flow directly to the adjacent contact, rather than having to flow down to the substrate, across, and up to a distant contact. Contacts placed immediately adjacent to a substrate injector can also help prevent localized debiasing of the highly resistive isolation, protecting adjacent tanks from injection from the isolation sidewalls.

Preventative Measures

Integrated circuits should inject as little current into the substrate as possible, as this not only minimizes substrate debiasing but also helps limit noise and cross-talk caused by modulation of the substrate potential. The collector current of substrate PNP transistors flows directly into the substrate, so these devices should be used sparingly, and no single device should conduct more than a milliamp or two. Lateral PNP and vertical NPN transistors can inject large substrate currents when they saturate, but techniques have been developed to minimize this problem (Sections 8.1.4–5). The exact requirements for substrate contacts depend on the nature of the substrate and isolation:

Heavily doped substrates. The contacts in the scribe seal can usually extract 5 to 10mA without undue debiasing. If higher substrate currents are anticipated, then the total area of contacts required can be computed using the following formula:[34]

$$A_c = 10\frac{\rho t_{epi} I_s}{V_d} \qquad [4.1]$$

This formula assumes a uniform lightly doped isolation, such as the P-epi of N-well CMOS and BiCMOS processes. A_c represents the required total area of substrate contacts in μm², ρ is the resistivity of the epi in Ω·cm, t_{epi} is the epi thickness in microns, I_s is the maximum substrate current in milliamps, and V_d is the maximum allowable debiasing in volts (from Table 4.1). The thickness of the epi is reduced by up-diffusion of dopants from the underlying substrate and from the presence of a heavily doped (if thin) contacting diffusion, such as PSD. Consider a die with a P-epi resistivity of 10Ω·cm and an effective epi thickness of 7μm. If the substrate must conduct 20mA without more than 0.3V of debiasing, then 47,000μm² (72mil²) of substrate contacts are required. Subtracting the area of substrate contacts in the scribe seal yields the required area of additional contacts. These can be inserted wherever space exists in the layout. As a precaution against localized debiasing, substrate contacts should ring any device injecting more than 1mA.

[34] This formula is derived from the fundamental equation $R = \rho l/A$. It neglects fringing effects, which tend to reduce the effective resistance of small substrate contacts. The formula provides a first-order approximation of the worst-case substrate resistance.

Lightly doped substrates with heavily doped isolation. No simple formula exists for computing the area of substrate contacts required to protect a lightly doped substrate from debiasing. A scattering of ten or twenty substrate contacts across the die will, when combined with the scribe seal, handle at least 5 to 10mA. Any device that injects 100μA or more should have substrate contacts located nearby, and any device that injects 1mA or more should be ringed with as much substrate contact as possible. Sensitive low-current circuitry should reside at least 250μm (10mil) away from any substantial source of substrate injection, since debiasing on lightly doped substrates tends to localize around the point of injection. Once the layout has been completed, additional substrate contacts should be scattered throughout the layout wherever room exists. A large number of small substrate contacts scattered throughout the layout will prove more effective than a few large contacts. Even with all of these precautions, designs that inject more than 10mA into the substrate may experience debiasing. Apart from adding more substrate contacts or moving sensitive circuits away from substrate injectors, the only remedies for such problems are the addition of a heavily doped substrate or the use of backside contacting.

Lightly doped substrates with lightly doped isolation. A few processes use a lightly doped substrate in combination with a very resistive isolation. This situation can arise when a BiCMOS design is constructed on a lightly doped substrate to save costs. Such designs cannot rely on the scribe seal to extract more than a few milliamps of substrate current. Large numbers of substrate contacts scattered across the die will help extract substrate current, but some degree of localized substrate debiasing is almost inevitable. Sensitive circuits should be located far away from major sources of substrate injection. Since substrate modulation can inject substantial noise into high-impedance circuitry, consider placing wells under resistors and capacitors to isolate them from substrate noise coupling. Sensitive MOS circuitry may also employ NBL to isolate NMOS transistors from the substrate (Section 11.2.2). In some cases, it may be possible to add strips of heavily doped material to the isolation without increasing the well-to-well spacings (Figure 4.16). This strategy effectively converts the design into one that uses a lightly doped substrate in conjunction with a heavily doped isolation. This stratagem substantially reduces the number and area of substrate contacts required to extract large substrate currents. Backside contact can also provide a large reduction in substrate resistance, but it is difficult to obtain Ohmic contact to a lightly doped substrate unless a backside diffusion is performed to increase the surface doping concentration.

4.4.2. Minority-Carrier Injection

Junction isolation relies on reverse-biased junctions to block unwanted current flow. The electric fields set up by depletion regions repel majority carriers, but they cannot block the flow of minority carriers. If any isolation junction forward-biases, it will inject minority carriers into the isolation. Many of these carriers recombine, but some eventually find their way to the depletion regions isolating other devices.

Effects

Figure 4.19 shows a cross section of a standard bipolar circuit. Suppose that the collector of NPN transistor Q_1 connects to a pin of the integrated circuit, and that the external circuitry experiences occasional transient disturbances that pull current out of this pin. If transistor Q_1 is off, then these transients pull its tank below ground, forward-biasing the collector-substrate junction of Q_1 and injecting minority carriers (electrons) into the substrate. Most of these carriers recombine, but some diffuse across to other tanks, such as T_1.

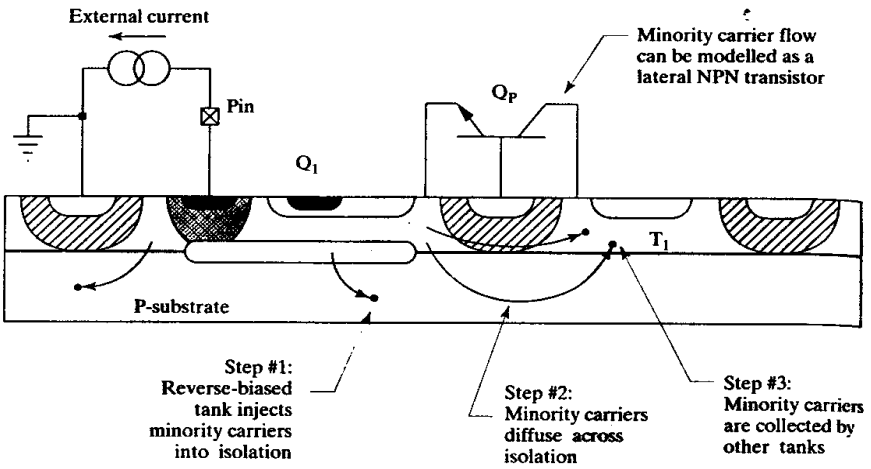

External current
Pin
Q_1
Q_P
Minority carrier flow can be modelled as a lateral NPN transistor
T_1
P-substrate
Step #1: Reverse-biased tank injects minority carriers into isolation
Step #2: Minority carriers diffuse across isolation
Step #3: Minority carriers are collected by other tanks

FIGURE 4.19 Example of minority-carrier injection into the substrate of a standard bipolar process. Lateral NPN transistor Q_P models the transit of minority-carriers across the isolation.

The transit of minority carriers across the isolation is analogous to the flow of minority carriers through a bipolar transistor. The tank pulled below ground acts as the emitter of lateral NPN transistor Q_P. The isolation and substrate act as the base of this transistor, and any other reverse-biased tank acts as a collector. Each reverse-biased tank forms a separate parasitic transistor corresponding to Q_P. The betas of these parasitic lateral NPN transistors are very low because most of the minority carriers recombine in transit. The parasitic bipolar between two adjacent tanks might have a beta of 10, but the beta between two widely separated tanks might not even reach 0.001. Even such low gains can cause circuit malfunctions. Suppose that a forward-biased tank injects a minority current of 10mA into the substrate. If the parasitic associated with another tank has a beta of 0.01, then this tank will collect 100μA of current—easily enough to disrupt the operation of a typical analog circuit.

Substrate contacts cannot, by themselves, stop minority-carrier injection since minority carriers travel by diffusion and not by drift. Minority carriers are best collected by reverse-biased junctions. However, substrate contacts still provide majority carriers to feed recombination. Since most minority carriers recombine in the isolation, substrate contacts remain necessary to prevent substrate debiasing.

In some cases, minority-carrier injection can cause a circuit to latch up. Early CMOS processes suffered from a form of this malady that has since come to be called *CMOS latchup.*[35] Figure 4.20A shows the cross section of a portion of a CMOS die consisting of an NMOS transistor M_1 and a PMOS transistor M_2. In addition to these two desired MOS transistors, this layout contains two parasitic bipolar transistors. Lateral NPN transistor Q_N's emitter is the source of M_1, its base is the isolation, and its collector is the N-well of M_2. Lateral PNP transistor Q_P's emitter is the source of M_2, its base is the N-well, and its collector is the isolation. Figure 4.20B shows the two parasitic bipolar transistors drawn in a more familiar fashion. In this schematic, R_1 represents the well resistance of M_2, and R_2 represents the substrate resistance. These two resistors normally ensure that both bipolar transistors remain off. As long as this remains the case, neither parasitic conducts any current and the integrated circuit works as intended. When a transient disturbance turns on either transistor, the current flowing through this device will turn on the other parasitic as well. Each transistor then supplies the other's base current. Once both tran-

[35] R. R. Troutman, "Recent Developments in CMOS Latchup," *IEDM Tech. Dig.*, 1984, pp. 296–299.

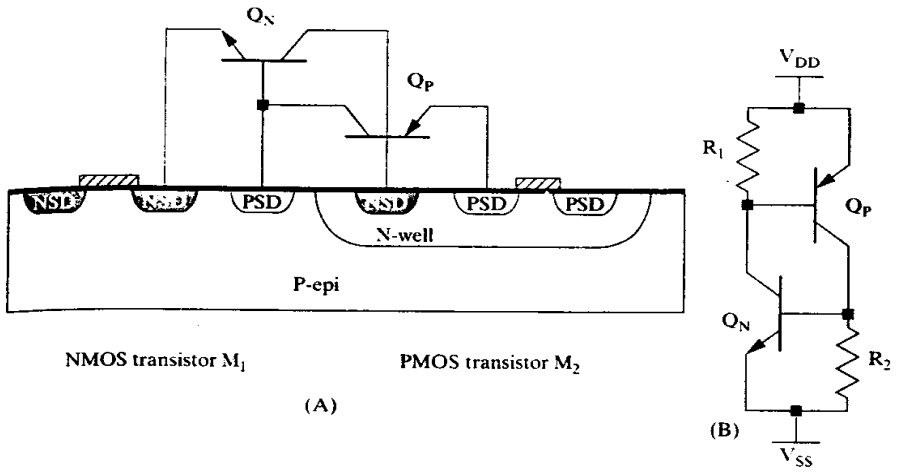


FIGURE 4.20 (A) Cross section of a CMOS die showing diffusions that form the two parasitic bipolar transistors Q_P and Q_N; (B) equivalent schematic showing these transistors along with well resistance R_1 and substrate resistance R_2.

sistors begin to conduct, they will continue to do so even if the transient disturbance that initiated conduction is removed. The circuit has *latched up* and it will remain in this state until power is removed. The integrated circuit can actually conduct so much current that it overheats and self-destructs. Even if this does not occur, latchup causes circuit malfunctions and excessive supply current consumption.

CMOS latchup can be triggered in one of two ways. If the source of NMOS transistor M_1 is pulled below ground, it will inject minority carriers (electrons) into the substrate, turning on parasitic transistor Q_N. This transistor will then turn on Q_P. Alternatively, the source of PMOS transistor M_2 may be pulled above the well. It will then inject minority carriers (holes) into the well and will turn on parasitic transistor Q_P. This transistor then turns on Q_N. Latchup can only occur if the product of the betas of transistors Q_N and Q_P exceeds unity. If the beta product is less than unity, then transient disturbances may occur, but the circuit cannot actually latch up (Section 11.2.7). CMOS latchup also requires the existence of four distinct semiconductor regions arranged in the sequence PNPN. A discrete device called a *silicon controlled rectifier* (SCR) has this same four-layer structure. CMOS latchup is sometimes described in terms of a parasitic SCR consisting of PSD, N-well, P-epi and NSD. This point of view is analogous to the transistor approach discussed above because the SCR operates in the same manner as the pair of coupled transistors.

The obvious way to stop CMOS latchup consists of reducing the beta of either or both parasitic transistors. If the product of these betas is less than unity, then latchup cannot occur. This is usually achieved by increasing layout spacings, which in turn increases the width of the neutral base regions of the parasitic lateral transistors. Alternatively, the amount of dopant present in the neutral base region of one (or both) parasitic transistors may be increased. Both of these approaches increase the Gummel number of one or both transistors and reduce the beta product.

Although many CMOS processes claim immunity to latchup, these claims are true only in a somewhat narrow sense. The PNPN structure inherent in the CMOS transistors of such a process lacks sufficient gain to establish regenerative feedback, but minority-carrier injection still occurs. The collected carriers can still cause circuit malfunctions, and if positive feedback exists in the circuit, these malfunctions can still cause a form of latchup. The significance of this observation is frequently underestimated. Any integrated circuit that experiences unanticipated minority-carrier injection can potentially latch up. Even if it does not actually do so, it is still likely

to malfunction. Not only do electrons injected into the substrate pose a potential threat, but so do holes unintentionally injected into wells or tanks.

Preventative Measures (Substrate Injection)

Fundamentally, there are four ways to defeat minority-carrier injection: (1) eliminate the forward-biased junctions that cause the problem, (2) increase the spacing between components, (3) increase doping concentrations, and (4) provide alternate collectors to remove unwanted minority carriers. All of these techniques provide some benefit, and in combination they can correct almost any minority-carrier injection problem.

The simplest solution, at least in theory, consists of eliminating the forward-biased junctions that inject minority carriers. This goal is often very difficult to achieve. In a standard bipolar process, tanks must not go below substrate by more than about 0.3V or they will inject minority carriers into the substrate. In an N-well CMOS process, no well and no NSD region residing in the epi may go below substrate potential. If the voltage on a pin slews rapidly, parasitic inductance can cause transients that pull the pin above supply or below ground. The faster the node slews, the smaller the parasitic inductance required to cause such transients. Modern switching speeds have become so fast that the inductance of pin and bondwire alone often cause objectionable transients. Substrate injection has become difficult, if not impossible, to eliminate.

Minority-carrier injection into the substrate will cause fewer problems if potential injectors are separated from sensitive circuitry. In most designs, only a few devices connect to pins. With a little forethought, these devices can be placed far away from sensitive circuitry. In many cases, the layout will naturally favor this sort of separation. For example, power transistors inject minority carriers during transients. Since these transistors form part of the output circuitry, they will typically be placed far away from sensitive input circuitry to minimize electrical and thermal feedback. This same arrangement also minimizes the circuit's vulnerability to minority-carrier injection.

Additional dopant added to the isolation regions of the die will reduce the gain of the parasitic lateral bipolar. CMOS and BiCMOS processes often employ P+ substrates for just this reason. All other factors being equal, a process incorporating a heavily doped substrate will provide greater immunity to electrical upsets than one that uses a lightly doped substrate. However, a heavily doped substrate cannot, by itself, prevent minority carriers from moving laterally through isolation regions separating adjacent tanks or wells. In order to obtain the full benefits of the heavily doped substrate, the process must use a heavily doped isolation, or the designer must add suitable guard rings.

The isolation doping can also be increased by adding a deep-P+ diffusion. Most CMOS processes do not include any suitable diffusion. Some BiCMOS processes include one for constructing certain components (such as DMOS transistors)—usually as part of a process extension. Standard bipolar processes sometimes offer a deep-P+ process extension for constructing high-current lateral PNP transistors. If a suitable diffusion exists, it can be placed in the isolation regions of the die to help increase the isolation doping. This technique can help offset the very light doping of the PBL portion of an up-down isolation system, and can minimize lateral conduction of minority carriers between adjacent tanks or wells.

Minority carriers are collected in disproportionate numbers by reverse-biased junctions near the point of injection. Not only do carriers have less distance to travel to reach a nearby junction, but the nearer junctions also block the flow of carriers to more distant ones. Designers can take advantage of this *shadow effect* to erect delib-

erate barriers to the flow of minority carriers by placing reverse-biased junctions between the point of injection and vulnerable diffusions. A reverse-biased junction used in this manner is called a *minority-carrier guard ring*. Figure 4.21 shows a typical layout for such a guard ring in a standard bipolar process. Tanks T_1 and T_2 connect to pins that may experience voltage transients. These are surrounded by a third tank, T_3, that collects a significant fraction of the minority carriers injected by T_1 and T_2. Tanks T_1 and T_2 connect to pins, so it is quite natural to place them along one side of the die or even in a corner (as shown in Figure 4.21). This not only minimizes the length of interconnecting leads but also eliminates the need for guard rings along two edges.

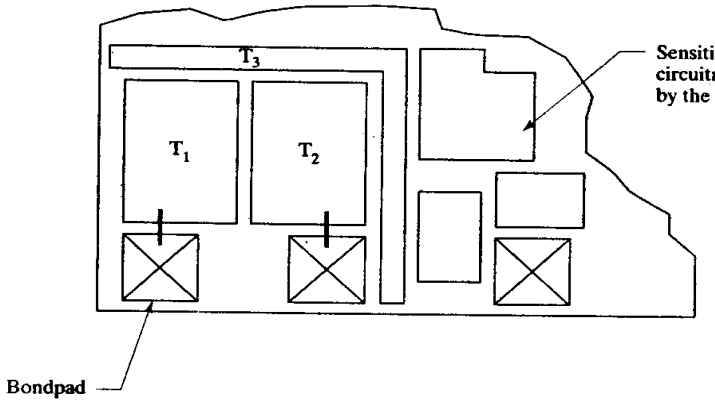


FIGURE 4.21 Sample electron-collecting minority-carrier guard ring (T_3) implemented in a standard bipolar process.

The key to designing efficient guard rings consists of making them deep, wide, and low-resistance. The deeper the guard ring, the larger the fraction of passing minority carriers it can collect. N-well makes a more effective electron-collecting guard ring than NSD, and an epi tank makes a better electron-collecting guard ring than an emitter diffusion. If the guard ring can be connected to produce a large reverse bias, then the depletion region surrounding it will widen and the collection surface will be forced even deeper into the silicon. Thus, electron-collecting guard rings connected to the positive supply become more effective than those connected to substrate potential. Since diffusing minority carriers move randomly, some will actually be collected by the bottom of the guard ring's depletion region. A wider guard ring will therefore collect more minority carriers than a narrow one will. Also, narrow diffusions do not penetrate as deeply as wide diffusions because dopants become diluted by lateral dispersion. A diffusion two or three times wider than minimum will contain enough dopant to obtain the maximum possible junction depth. Low resistance also helps improve the effectiveness of a guard ring, especially if it cannot be strongly reverse-biased. Collected carriers can forward-bias a high-resistance guard ring and cause it to re-inject minority carriers. The lower the vertical resistance of the guard ring, the larger the current it can collect before it saturates and re-injects.

Figure 4.22 shows cross sections of two minority-carrier guard rings designed to collect electrons injected into the substrate. Figure 4.22A shows a substrate guard ring for standard bipolar.[36] This guard ring includes all four available N-type materials: N-epi, deep-N+, NBL, and emitter. The NBL helps obtain the maximum possible junction depth, while deep-N+ and emitter reduce the vertical resistance. The wider this structure, the more effectively it will collect minority carriers. Most designers

[36] W. Davis. *Layout Considerations*, unpublished manuscript, 1981, p. 53.

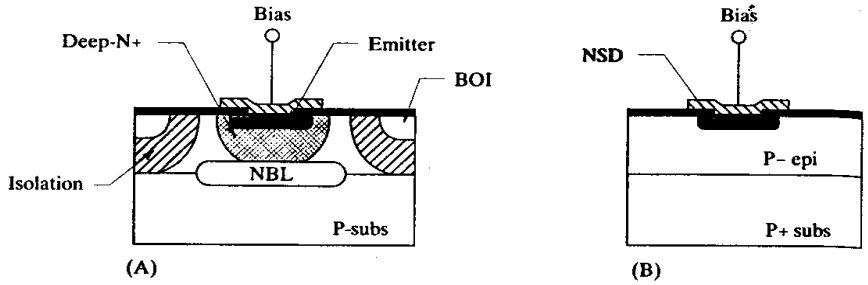


FIGURE 4.22 Cross sections of two representative electron-collecting guard rings: (A) standard bipolar[37] and (B) CMOS.

compromise between efficiency and area by making the deep-N+ strip no more than twice minimum width and by spacing the other layers accordingly. If possible, this guard ring should connect to the highest supply voltage available on the die. The guard ring will still function connected to substrate potential, but it may saturate unless all parts of the guard ring connect to the substrate terminal by a direct metal run. This is probably not practical in a single-level metal process since gaps must be left in the metallization to allow leads to pass through. Single-level-metal guard rings should connect to the positive supply and should contain as few gaps as possible.

BiCMOS layouts can produce electron-collecting guard rings similar to those in Figure 4.22A, although in this case N-well replaces the N-epi tank. Electron-collecting guard rings are considerably more difficult to construct in CMOS-only processes. The N-well becomes extremely resistive in the absence of deep-N+ and NBL, and most CMOS devices operate at relatively low voltages. Figure 4.22B shows an alternate CMOS minority-carrier guard ring. NSD has a relatively low resistance, but it is too shallow to capture more than a small percentage of the electrons in the substrate. The wider the NSD strip, the more effective the guard ring. A width of at least 8 to 10μm is recommended, although narrower guard rings do provide some benefit. The NSD guard ring should, if possible, connect to a supply pin. The reverse bias across the NSD-epi junction drives the depletion region deeper into the epi and increases the apparent depth of the guard ring. In low-voltage processes, a strongly reverse-biased NSD guard ring will often generate secondary carriers due to impact ionization. This problem can be minimized by connecting the low-voltage NSD guard ring to ground instead of to a power supply.

Guard rings of the type shown in Figure 4.22A can reduce substrate injection by a factor of 10 to 100 providing they are used in conjunction with a heavily doped substrate. The P–/P+ interface between the lightly doped epi and the heavily doped substrate repels minority carriers (Section 8.1.5), constraining them to remain within the relatively thin epi layer. This greatly improves the collection efficiency of the guard ring.[38] A simple modification to the electron-collecting guard ring of Figure 4.22A can further increase its attenuation. Instead of connecting the guard ring directly to a supply voltage, it is connected back to the substrate so that a majority-carrier current flows through the substrate beneath the guard ring (Figure 4.23). This type of guard ring is intended to protect against minority carriers originating on only one side of the ring—in this case, the right side. The deep-N+ sinker in the center of the guard ring collects most of these minority carriers. The resulting cur-

[37] C. Jones, "Bipolar Parasitics," unpublished report, 1988, p. 43.

[38] L. S. White, G. R. M. Rao, P. Linder, and M. Zivitz, "Improvement in MOS VLSI Device Characteristics Built on Epitaxial Silicon," in *Silicon Processing*, American Society for Testing and Materials STP 804, 1983, pp. 190–205.

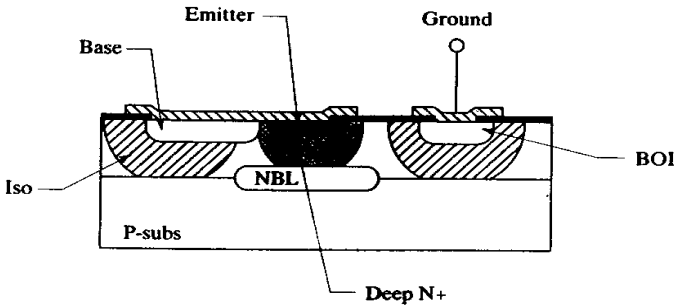


FIGURE 4.23 Cross section of an improved minority-carrier guard ring for collecting electrons injected into the substrate.

rent flows out of the sinker and into the attached metal plate. The base/iso diffusion at the left side of the structure re-injects this current into the substrate in the form of majority carriers. Since the nearest substrate contact lies on the other side of the structure, the majority carriers flow back underneath the guard ring. This current locally debiases the substrate and creates an electric field that opposes minority-carrier flow. The minority carriers are forced upward and toward the guard ring, where they are ultimately collected, or they are held in the substrate until they recombine. The inventor[39] claims an attenuation factor in excess of one million for this structure. While this degree of attenuation may not be achieved in every process, this guard ring will provide more attenuation than those shown in Figure 4.22, especially at higher currents.

The modified guard ring in Figure 4.23 suffers from several drawbacks. It provides enhanced attenuation of carriers flowing in only one direction—in this case from right to left. Substrate contacts cannot be placed near the guard ring on the side facing the injected carriers. This structure also relies upon deliberate debiasing of the substrate, which could potentially forward-bias adjacent junctions. The principle behind this style of guard ring also applies to the design of ordinary guard rings of the sort shown in Figure 4.22. In all cases, it is better to place the electron-collecting guard ring adjacent to the injector, and to place substrate contacts inside of the guard ring. The majority-carrier substrate current that flows under the guard ring generates an electric field opposing the flow of minority carriers, thereby enhancing the performance of the guard ring.

Minority-carrier guard rings can also prevent holes injected into a tank from reaching the substrate and debiasing it. This situation can occur whenever a bipolar transistor saturates, regardless of whether this transistor is an NPN or a PNP (Section 8.1.4–5). Power transistors can easily inject tens or even hundreds of milliamps into the substrate. A heavily doped layer such as NBL can reduce minority-carrier injection from a P-type region through an N-well or N-epi tank to substrate. NBL also helps to minimize tank or well resistance and therefore makes it more difficult to develop the debiasing required to trigger CMOS latchup. CMOS processes generally do not incorporate NBL due to the cost of the extra masking step and to manufacturing difficulties associated with the fabrication of buried layers. Standard bipolar and analog BiCMOS processes frequently use NBL to reduce NPN collector resistance. If NBL exists, it should be added to all tanks or wells that can tolerate its presence, in order to minimize substrate injection and to improve latchup immunity.

[39] F. Van Zanten, U.S. Patent # 4,466,011, 1984.

Figure 4.24 shows a hole-collecting guard ring constructed in a standard bipolar process. As a first line of defense against hole injection into the substrate, the tank is floored with NBL and ringed with deep-N+. Any hole attempting to reach the isolation must pass through one or the other of these heavily doped regions. The large population of electrons in these regions enhances recombination and prevents most holes from successfully crossing. Far more importantly, the N+/N– boundary exhibits a built-in potential gradient caused by the outdiffusion of majority carriers that helps confine holes inside the tank until they recombine or are collected (Section 8.1.5).

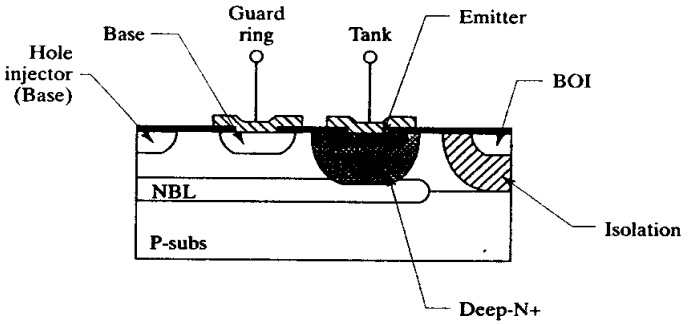


FIGURE 4.24 Cross section of a minority-carrier guard ring for collecting holes injected into a tank.[40]

The guard ring in Figure 4.24 includes a base ring placed just inside the deep N+. This ring usually connects to ground, but it will remain reasonably effective even if it is tied to the tank terminal. Any hole impinging on the depletion region surrounding the base ring will be drawn across by the electric field. Holes become majority carriers inside the base diffusion and can be removed through the contact. The base ring collects almost all of the holes as long as the tank contains NBL. Even without the deep-N+ ring around the outside edge of the tank, the base ring will collect at least 90% of the holes. This type of guard ring is largely ineffectual without NBL.

CMOS devices may experience hole injection into wells if a PSD region rises above the well potential, as might occur if a pin connected to a PMOS source or drain rises above supply. Effective hole collection rings cannot be constructed in a pure CMOS process due to the absence of NBL. The doping gradient of the well causes a downward drift of holes toward the underlying substrate, rendering PSD guard rings placed around the edges of the well ineffectual. CMOS processes must therefore rely upon low-resistance substrate contacts to extract any hole current injected into the substrate.

BiCMOS processes can construct hole-collection rings similar to those in Figure 4.24. These rings are not quite as effective on BiCMOS processes as on standard bipolar because the graded profile of the BiCMOS well opposes the potential barrier raised by the NBL. This problem is exacerbated by the relatively light doping of BiCMOS NBL regions required to avoid autodoping the P-epi. Despite these problems, an overall efficiency in excess of 95% is usually achievable by using base hole-collection rings in combination with NBL and deep N+. Section 13.2 discusses several additional types of hole guard rings and some of the difficulties associated with constructing hole guard rings in a BiCMOS process.

[40] Jones, p. 18ff.

Preventative Measures (Cross-injection)

Circuit upsets caused by minority-carrier injection into a tank or well can often be eliminated by placing each potential emitter of minority carriers in its own tank or well. As a rule, any PMOS transistor whose source connects to an external pin should occupy its own well. Similarly, any base resistor, HSR resistor, or lateral PNP collector connecting to a pin is best placed in its own tank. The small amount of extra space required to construct separate tanks or wells will be amply repaid by the elimination of even one parasitic. If, on the other hand, several devices connect to a common pin, then these can all occupy a common tank or well.

The hole-collection rings discussed previously have been designed to minimize injection of holes into the substrate. Another type of minority-carrier guard ring can prevent holes injected by one device from interfering with the operation of other devices in the same tank or well, a problem called *cross-injection.* Consider the case where two lateral PNP transistors occupy a common tank. If either transistor saturates, some fraction of the carriers it emits will be collected by the adjacent transistor. The resulting increase in collector current may disturb the operation of the circuit, particularly if the devices were intended to match one another. Cross-injection can be prevented by placing each transistor in its own tank, but this wastes area because of the large spacings associated with the isolation diffusion. A more compact solution employs a type of minority carrier guard ring called a *P-bar* (Figure 4.25).[41]

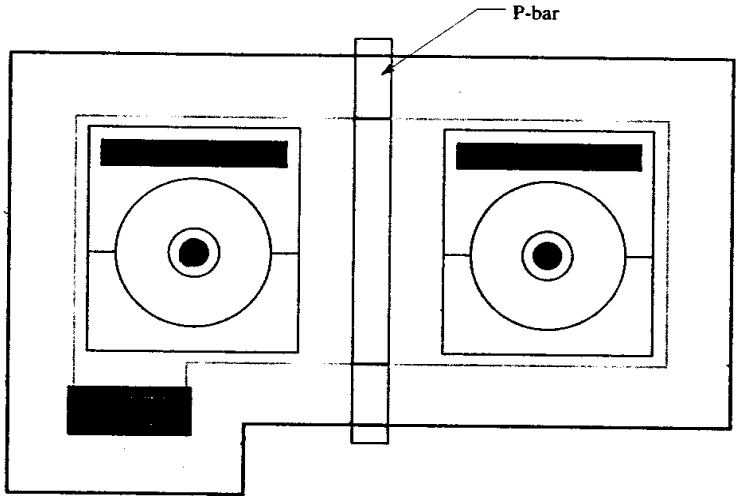


FIGURE 4.25 Example of a P-bar used to prevent cross-injection between two lateral PNPs.

A P-bar consists of a minimum-width strip of base diffusion placed between the two transistors.[42] Each end of the P-bar extends out into the isolation far enough to guarantee electrical contact. This arrangement ensures that the P-bar electrically connects to the isolation without requiring contacts. Now suppose that the lateral PNP on the left side of the P-bar saturates and begins to inject holes into the tank. In order for these holes to reach the lateral PNP transistor on the right, they must first pass underneath the P-bar. The base diffusion forming the bar reaches fairly deeply into the epi and leaves little room for carriers to pass underneath it. Most of

[41] Davis, p. 27; Jones, p. 10.
[42] Jones, p. 10.

the holes traveling from left to right will be collected by the P-bar and shunted to ground. This structure thus acts as a specialized type of minority-carrier guard ring. The presence of NBL beneath the P-bar ensures a low-impedance path for base current passing from the right-hand transistor to the tank contact at the lower left corner of the tank. The tank contact on the left side of this structure therefore suffices for both transistors.

Although the collection efficiency of P-bars can only be determined through empirical measurement, several observations are in order. As the tank bias increases, the depletion region surrounding the bar deepens and progressively pinches off the N-epi underneath it. Devices operating at high tank-to-substrate potentials therefore obtain a higher degree of isolation from a P-bar than devices operating near substrate potential. A wider P-bar also increases collection efficiency, in part because the pinched portion of the tank becomes wider and in part because the wider base region diffuses deeper into the epi. Even a minimum-width P-bar provides a high degree of isolation against minority-carrier cross-injection due to the up-diffusion of the underlying NBL and the formation of a depletion region beneath the P-bar.

The P-bar has many applications. Bipolar circuits often contain current mirrors composed of lateral PNP transistors with a common base connection. These transistors often occupy a common tank, but if one transistor saturates then the currents provided by the adjacent transistors increase. P-bars placed between the saturating transistor and the adjacent devices will prevent this effect without unduly enlarging the tank. Another common application consists of an NPN driving either a lateral or a substrate PNP transistor, in which the collector of the NPN connects to the base of the PNP. Minority-carrier conduction from the PNP to the NPN can initiate positive-feedback latchup by triggering the SCR inherent in this structure. A P-bar placed between the transistors may suppress the latchup, although this is not guaranteed unless the collection efficiency of the bar exceeds the reciprocal of the beta product of the two transistors.

P-bars also find use in CMOS processes, where they typically consist of PMoat. This type of P-bar exhibits a lower collection efficiency than its bipolar counterpart due to the shallowness of the PMoat diffusion and the absence of NBL. The lack of a buried layer greatly increases the well resistance beneath the P-bar, so prudence dictates the inclusion of well contacts on both sides of the bar. This structure can help increase a circuit's latchup immunity without requiring separate wells. If one PMOS transistor in the tank has a source or drain connecting to an outside terminal, then a transient can potentially forward-bias this PMoat into the well. The resulting minority-carrier injection can disturb adjacent transistors and can even lead to latchup. The strategic placement of a few minimum-width PMoat P-bars can provide considerable protection against this sort of cross-injection without consuming as much area as separate wells require.

Another type of minority-carrier guard ring called an *N-bar* can also protect against minority-carrier cross-injection. An N-bar consists of a strip of deep-N+ placed between two devices occupying a common tank (Figure 4.26).[43] The N-bar typically serves as a tank contact for the devices around it since the spacings surrounding the deep-N+ are large enough to allow room for both emitter diffusion and a contact. The doping gradient surrounding the N-bar repels minority carriers, and most of the carriers that overcome this gradient recombine inside the deep-N+ before they pass through it. Unfortunately, the N-bar generally stops short of the

[43] Davis, p. 31.

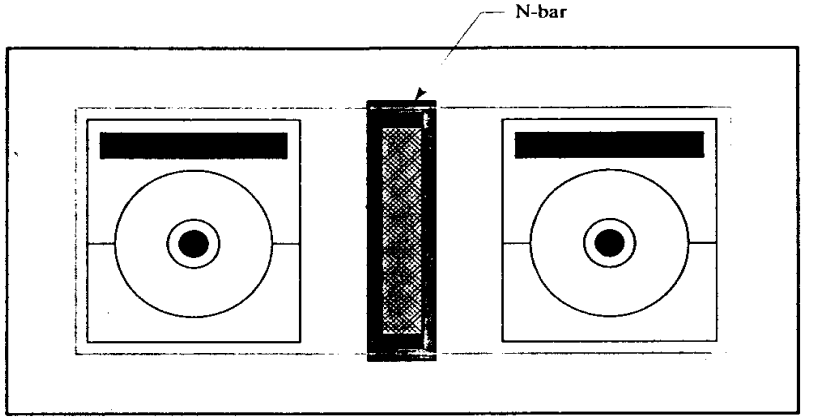


FIGURE 4.26 Example of an N-bar used to simultaneously provide tank contact and to minimize cross-injection between two lateral PNP transistors.

P-isolation on either side of the tank to avoid forming an N+/P+ junction that would break down at a relatively low voltage. These gaps allow minority carriers to bypass the N-bar, so an N-bar usually exhibits a lower collection efficiency than a P-bar. Still, the combination of a highly efficient collector contact with a moderately effective minority-carrier guard ring sometimes finds applications in high-current lateral PNP current mirrors and similar circuits.

4.5 SUMMARY

This chapter discusses a number of common failure mechanisms of integrated circuits. Table 4.2 summarizes these mechanisms, along with typical symptoms and suggested corrective actions. Even a cursory glance at the table reveals the interdisciplinary nature of the subject. Some mechanisms are primarily electrical, while others depend upon chemical or electrochemical processes. Some of these failure mechanisms require a knowledge of device physics to counteract them, while others require knowledge of processing and packaging technology. Only by amassing a working knowledge of many fields can one hope to design integrated circuits that will function reliably over a lifetime of use.

4.6 EXERCISES

Refer to Appendix C for layout rules and process specifications.

4.1. A certain copper-doped aluminum alloy can safely operate at current densities of $5 \cdot 10^5 A/cm^2$. If the metallization thickness equals 8kÅ, but thins by 50% when passing over oxide steps, then how much current can a 10μm-wide lead carry across an oxide step?

4.2. Propose a scribe seal structure for a single-level-metal standard bipolar process. Draw a cross section of this structure and explain the purpose of each of its components.

4.3. Lay out a 15kΩ, 8μm-wide HSR resistor. Field plate the resistor as well as possible, including flanges where necessary. The field plate should overhang HSR by at least 6μm and base by at least 8μm.

4.4. Modify the layout from Exercise 4.3 to include channel stops constructed from emitter diffusion. Assume that the channel stops must overlap the field plates by 4μm.

4.5. Construct a minimum-size, standard-bipolar, lateral PNP using a circular emitter geometry. Fully field-plate both the emitter and the collector, leaving space for base metallization. Assume the emitter field plate must overlap the collector by 2μm and the collector field plate must overhang the collector by 8μm.

TABLE 4.2 Summary of failure mechanisms.

Failure Mechanism	Symptoms	Corrective Actions*
Electrostatic discharge (ESD)	Gate oxide ruptures either immediately or after delay, junctions shorted or leaky.	***Add ESD protection devices, do not route leads over thin emitter oxide.***
Electromigration	Open or short circuits after long-term operation, usually at high temperature.	Use copper-doped aluminum, use refractory barrier metal, ***use adequate lead widths, use adequate bondwires.***
Antenna effect	Small gate oxides connected to large conductors suffer delayed failure.	***Reduce ratio of conductor area to gate oxide area, add diodes.***
Dry corrosion	Open circuit failures, moisture accelerates failure.	Use nitride PO, ***minimize PO openings.***
Mobile ions	Threshold shifts under high-temp biased operation, relaxes after unbiased bake.	Use phosphosilicate glass, use polysilicon gate MOS, ***minimize PO openings, use adequate scribe seals.***
Hot carriers (in MOS)	Threshold shifts under high-temp biased operation, relaxes after unbiased bake.	***Limit drain-source voltages, use LDD structures, use long-channel devices.***
Zener walkout	Breakdown voltage drifts, relaxes after unbiased bake.	***Use buried Zener*** (if available).
Parasitic channels & Charge spreading	Leakage currents at high voltage. If they appear after high-temp biased operation and relax after high-temp bake, charge spreading is responsible.	Use (111) silicon, add channel stop implants, ***add base-over-iso, use channel stops, use field plates.***
Substrate debiasing	Latchup, parametric shifts that occur under specific biasing conditions.	***Maximize substrate contact, place contacts near injectors.***
Minority-carrier injection into the substrate	Latchup, parametric shifts that occur under specific biasing conditions.	Use P+ substrate, ***maximize substrate contact, separate sensitive circuitry, add NBL to shared wells, use deep-P+ in isolation, add guard rings.***
Minority carrier cross-injection	Latchup, mismatches between merged devices.	***Use P-bar or N-bar, place devices in separate tanks or wells.***

* Possible solutions listed in ***bold italics*** are under the control of circuit and layout designers; the remaining solutions can only be implemented by process engineers.

4.6. Compute the area of substrate contacts necessary to extract 25mA from a die that uses an 8μm-thick, 10Ω·cm, P-type epi layer on top of a 0.01Ω·cm P-type substrate. Assume a maximum allowed debiasing of 0.3V.

4.7. Lay out a standard-bipolar NPN transistor with a 20μm by 40μm emitter. Arrange the transistor to minimize the distance between the emitter and collector contacts. The transistor should include deep-N+ in the collector to reduce collector resistance. Place an electron-collecting guard ring around this transistor, following the cross section shown in Figure 4.22.

4.8. Lay out a 2000/5 PMOS transistor. Divide the transistor into a sufficient number of fingers to obtain a roughly square aspect ratio. Construct a hole-collecting guard ring similar to that in Figure 4.24 that encircles the PMOS transistor. Make sure that the NBL overlaps the deep-N+ diffusion by at least 4μm to provide an adequate seal at the point of intersection.

4.9. Lay out an example of a P-bar separating two minimum-size standard bipolar lateral PNP transistors. The P-bar should extend at least 4μm into the isolation to ensure electrical continuity.

4.10. Several failed devices have been de-encapsulated (*decapped*) for microscopic examination. Suggest at least one failure mechanism consistent with each of the following observations:

a. A metal trace from a bond pad has melted open.
b. A greenish deposit covers the bond pads.
c. The gate oxide of a minimum-size NMOS has ruptured at one point, shorting the poly gate to the underlying epi.
d. A thin, dark filament appears across the base of a large NPN transistor. The transistor's base-collector junction appears shorted.

4.11. A new high-voltage, low-current operational amplifier has just completed burn-in testing. Sample units were operated under bias at 150°C for 1000 hours. Parametric testing reveals that the input offset voltages of the amplifier have drifted several millivolts during testing, and the supply currents have increased by 20%. What failure mechanisms might be responsible for these symptoms, and how can the designer determine what to fix?